UNCERTAINTY QUANTIFICATION OF GAS PRODUCTION IN THE BARNETT SHALE USING TIME SERIES ANALYSIS

By

Kishan Joshi

RECOMMENDED:

Dr. Obadare Awoleke
Advisory Committee Chair

Dr. Mohabbat Ahmadi
Advisory Committee Member

Dr. Catherine Hanks
Advisory Committee Member

Dr. Abhijit Dandekar,
Chair, Department of Petroleum Engineering

APPROVED:

Dr. Douglas Goering
Dean, College of Engineering and Mines

Dr. John Eichelberger
Dean of the Graduate School

12/17/15
Date

UNCERTAINTY QUANTIFICATION OF GAS PRODUCTION IN THE BARNETT SHALE USING TIME SERIES ANALYSIS

A

THESIS

Presented to the Faculty

of the University of Alaska Fairbanks

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

By

Kishan Joshi, B.Tech.

Fairbanks, Alaska

December 2015

Abstract

Deterministic methods for evaluating uncertainty in production forecasts for unconventional shale plays are either unreliable or time intensive. This thesis presents an improved methodology for quantifying uncertainty in production forecasts using Logistic Growth Analysis (LGA) and time series modeling. The applicability of the proposed method is tested by history matching production data and providing uncertainty bounds for forecasts from eight Barnett Shale counties. The 80% confidence interval (CI) generated by this method successfully bracketed true production values for all the counties, even when approximately one-third of the data was used for history matching.

In the methodology presented, the trend in the production data was determined using two different non-linear regression schemes. The predicted trends were subtracted from the actual production data to generate two sets of stationary residual time series. Time series analysis techniques (Auto Regressive Moving Average models) were thereafter used to model and forecast residuals. These residual forecasts were incorporated with trend forecasts to generate our final 80% CI.

To check the reliability of the proposed method, I tested it on 100 gas wells with at least 100 months of available production data. The CIs generated covered true production 84% and 92% of the time when 40 and 60 months of production data were used for history matching, respectively. An auto-regressive model of lag 1 best fit the residual time series in each case.

The proposed methodology is an efficient way to generate production forecasts and to reliably estimate uncertainty for short to medium time periods. It includes uncertainty due to parameter estimation using two different regression schemes. It also incorporates the uncertainty due to the variance of the residuals. The method is computationally inexpensive and easy to implement. The utility of the procedure presented is not limited to gas wells; it can be applied to any type of well or group of related wells.

Dedicated to

My beloved parents for their endless love and support.

The late Larsika Kropi, who was one of the kindest souls I've encountered, and a loving friend.

And my dear friends for their constant encouragement.

Table of Contents

List of Figures

Page

List of Tables

Page

List of Appendices

Page

Acknowledgments

I would like to express my sincere gratitude to my principal advisor, Dr. Obadare Awoleke, for his constant encouragement, support, and invaluable guidance during my research. I would also like to thank my advisory committee members, Dr. Mohabbat Ahmadi and Dr. Catherine Hanks, for their valuable suggestions and commitment to this thesis. I am also thankful to Dr. Ron Barry for providing useful discussions concerning the statistics side of my research.

I am grateful to the Chair of the Petroleum Engineering Department, Dr. Abhijit Dandekar, the Dean of the Graduate School, Dr. John Eichelberger, and the Alaska Department of Natural Resources for funding my research and graduate studies.

I express my deepest gratitude to my parents, Mr. Ghanshyam Joshi and Mrs. Pratima Joshi, for their unconditional love and support. Finally, I would like to thank my friends and Petroleum faculty for their support during my time at University of Alaska Fairbanks.

CHAPTER 1 INTRODUCTION

1.1 Production Forecasting for Unconventional Reservoirs

The term conventional reservoirs can be applied to hydrocarbon accumulations that can readily be extracted, after drilling operations, by natural pressure of the wells and pumping or compression methods (non-specialized methods). Conversely, unconventional reservoirs (typically tight sands, heavy oil, gas hydrates, and shale reservoirs) refers to hydrocarbon accumulations that will require new enhanced extraction or recovery techniques to produce economically viable amounts of hydrocarbons. In shale reservoirs, the most popular stimulation technique is hydraulic fracturing, which creates high conductivity pathways for hydrocarbons between the reservoir and the well. Such stimulation techniques are expensive and demand assurance of a certain amount of producible hydrocarbons to make them economically viable. Therefore, it is important to forecast production and estimate reserves from such reservoirs to aid project management and field development planning.

The complexity of production data, combined with limited analytical insights (geological and petrophysical properties) for unconventional reservoirs, makes estimating reserves and predicting production a challenge. This might be due to differences between the theory or assumptions of the models used to predict production (based on history matching) and actual observations in nature. For example, Arps' Decline Curve Analysis (DCA) assumes boundary dominated flow for the well under analysis, but in reality, it is difficult to predict the type of flow regime in unconventional wells. In a hydraulically fractured unconventional reservoir, production is observed from both the Stimulated Reservoir Volume (SRV) and from matrix permeability, the contribution to total production from each varying over time. Because of this difference between theory and reality, there is significant uncertainty in the deterministic estimates for unconventional reservoirs when traditional forecasting techniques are used. As a result, in a lot of cases, the deterministic future estimates of production are nowhere near observed production.

1.2 Non-deterministic Features in Prediction Models

Realistic modeling of production from unconventional reservoir systems must include non-deterministic features. Non-deterministic indicates that the response of the system (in our case, a

production profile) cannot be predicted accurately because (1) we do not have sufficiently accurate estimates of all the variables required to describe the system and (2) we do not have perfect models. Another source of uncertainty, arguably lesser (because these events are planned for), is human interaction with the system (Oberkampf et al., 2002). Human interaction may consist of, but is not limited to, well shut-ins, tertiary recovery techniques, and re-fracturing in shale reservoirs. Other sources of uncertainty include changes in the reservoir behavior over time, e.g., decreasing fracture conductivity as a function of depletion. It is our hope that engineers include as many mechanisms that drive well/reservoir behavior as possible and applicable. However, given that even under the best circumstances there is often significant uncertainty with production forecasts, it is crucial to quantify uncertainty in these forecasts to improve decision-making. My work attempts to quantify the uncertainty due to mechanisms we do not know about or can not estimate, or those that are intrinsic to the system under study.

In unconventional reservoirs, the use of decline curve analysis models is common. Typically, these deterministic modeling techniques provide a single matched model after history matching available production data (Arps, 1945; Clark et al., 2011; Duong, 2011; Ilk et al., 2008). For the reasons discussed above, it is incorrect to assume that a single model without any uncertainty quantification is sufficient for forecasting production in these wells. This emphasizes the need to incorporate non-deterministic elements in production forecast models in order to provide reliable confidence bounds given the available production data.

1.3 Objective and Approach

The fiscal attractiveness or feasibility of any field or well development project depends upon its ability to generate profit. This, in turn, depends on costs (both fixed and running) and temporal values of production data. Since engineers do not have production data a priori, analytical/empirical models are used to forecast production. Accordingly, for a robust economic analysis, the forecasts should include uncertainty estimates.

In this thesis, the applicability of time series analysis to quantify uncertainty in production from gas wells in the Barnett Shale was tested. The data was de-trended first using the Logistic Growth Analysis (LGA) model. De-trending refers to subtracting the deterministic (trend) component from the production data. Integrating both time series and LGA modeling enabled the author to generate

confidence intervals around the production forecast as reliably as Modified Bootstrap Method (MBM). First, the advantages of using the LGA model are discussed. Then uncertainty is quantified by modeling the resulting residual using time series analysis. Finally, the results of this analysis on a county-by-county basis for eight Barnett Shale counties are presented, and on a well-by-well basis for 100 Barnett Shale wells with at least 7 years of production data.

CHAPTER 2 DETERMINISTIC MODELS

2.1 Justification for Choice of Deterministic Model

Arps' DCA (1945) is the most popular tool for forecasting oil or gas production and reserve estimation in both conventional and unconventional reservoirs. The decline curve model assumes boundary-dominated flow for the well under study, with the value of the decline exponent b between zero and one. However, in low permeability shale gas plays, transient flow dominates the flow regime instead of boundary-dominated flow, resulting in a b value greater than one (Valko and Lee, 2010). This results in infinite cumulative production as time approaches infinity. Consequently, in practical terms, Arps' model usually overestimates reserves (Clark et al., 2011; Ilk et al., 2010) in unconventional reservoirs.

On the other hand, the logistic growth model does not extrapolate to non-physical values (Clark et al., 2011). Recent work by Paryani and others (in review) indicates the logistic growth model outperformed Arps' model when both models were used to match production history from 100 oil wells in the Eagle Ford Shale. The LGA model matched the wells' past production history 81% of the time, whereas Arps' DCA matched it less than 40% of the time. Based on this observation, we chose to use the LGA model for this study over Arps' model and its variants. The reader should note that the objective of this work is not to ascertain which deterministic model is the best, but to develop a methodology with which we can quantify uncertainty in production forecasting. As long as the data is properly de-trended and the deterministic model does not result in non-physical results (e.g., infinite reserves), any decline curve analysis or deterministic model can be used with this methodology.

2.2 Logistic Growth Analysis (LGA) Model

Logistic growth curves are a set of models used to forecast growth in various disciplines. Originally developed by Verhulst (1838), logistic growth models were used to model population and yeast growth, and for the study of liver regeneration. Spencer and Coulombe (1966) modified the generalized form of the logistic growth model to represent liver regrowth, which follows a hyperbolic trend. Clark and others (2011) adapted this form of the LGA model and modified it to

represent hyperbolic decline of production from extremely low permeability oil and gas wells. The modified form of the logistic growth model presented by Clark and others (2011) is shown below.

$$Q(t) = \frac{Kt^n}{a + t^n} \tag{1}$$

In **Eq. 1**, Q refers to cumulative production, which is analogous to regenerated liver size at a particular time in the liver regrowth experiment. K represents carrying capacity in logistic growth models, which is the threshold value a dependent variable under study may physically reach or grow to, at which point the variable will stabilize and cease to regenerate or grow. In Eq. 1, K refers to Estimated Ultimate Recovery (EUR), which is the potentially recoverable quantity of oil and gas from primary drive mechanisms in the well, considered without time or economic constraints. In the liver regrowth experiment (Spencer and Coulombe, 1966), K refers to the original size of liver before it was reduced for regeneration analysis. a is a constant, n is the hyperbolic exponent, and t is time elapsed during the process.

Eq. 1 is for forecasting cumulative oil or gas production with variation in time. If we take the derivative of Eq. 1 with respect to time, t, we obtain the equation for forecasting monthly production or production per unit time, **Eq. 2:**

$$q(t) = \frac{dQ}{dt} = \frac{Knat^{n-1}}{(a + t^n)^2} \tag{2}$$

Where $q(t)$ refers to monthly production or production per unit time.

2.3 Overview of Time Series Analysis

A dictionary definition of time series is a sequence of measurements of a quantity obtained at successive times or in a successive order, usually occurring at regular intervals. The sampling of adjacent points in time leads the time series to have internal structures, like autocorrelation, trend, and/or seasonal variation, all of which need to be accounted for. Time series analysis methods analyze time series data to extract its statistical characteristics. The scientific application of time series analysis to various disciplines is well documented (Shumway and Stoffer, 2011). The monthly gas production from a well is the time series that will be analyzed in this work.

A time series can be considered a vector, $X(t)$, of n>1 observations, x_i, observed at points in time, t_i, for i=1, 2…, n. Such a time series may be decomposed into its components, that is, its trend, seasonal and random components as shown in **Eq. 3** below.

$$X(t) = T(t) + S(t) + R(t) \tag{3}$$

Where: $X(t)$ = value of series at time t—for example, gas production at a time t; $T(t)$ = trend component of the series: results from LGA model at time t; $S(t)$ = cyclical or seasonal component (with specified period S), usually absent in gas production data; and $R(t)$ = white noise or random effect for which we have no explanation.

Since the noise component or $R(t)$ is the part of the data we can not explain with the deterministic model, I will estimate it using the trend and seasonality terms, when applicable. The trend is the long term change in mean and variance of the series and can be linear or nonlinear. The seasonal component represents a repeated or systematic pattern in a data series over a known fixed interval of time. The seasonality term will be ignored in this work.

For the purpose of quantifying uncertainty using time series analysis, it is important that the series under study is stationary, that is, it has constant mean and variance (no trend or seasonality). To the best of our knowledge, there is no seasonal component associated with monthly gas production data for a well drilled in the Barnett Shale. This implies that we only need to estimate the trend component, which can be achieved using the LGA model and nonlinear least squares regression.

"Reduction to noise" or de-trending, in our case, refers to reducing the time series to stationary white noise series or residuals through the process of subtracting the predicted production profile estimated using LGA and observed data (from the actual production data), as shown by **Figs. 2.1** and **2.2**. Fig. 2.1 is a semi-log plot of production data with the LGA trend. The resulting noise or residual is approximately stationary. Its components are also correlated with each other in some manner. Time series analysis methodology will be used to discover the mathematical nature of this correlation; this is called the error or residual model. This estimated residual model will be used to forecast residuals with uncertainty bounds. The residual forecast will then be incorporated with the trend forecast to generate a confidence interval around our production forecasts.

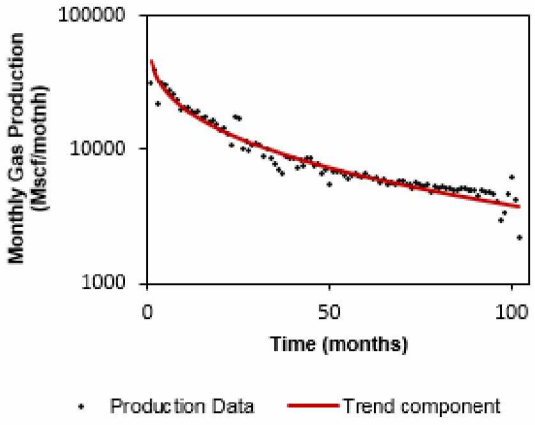

Figure 2.1: Monthly production profile generated by LGA estimated model for Well 1 using nonlinear least square regression (semi-log) (source: DI Desktop)

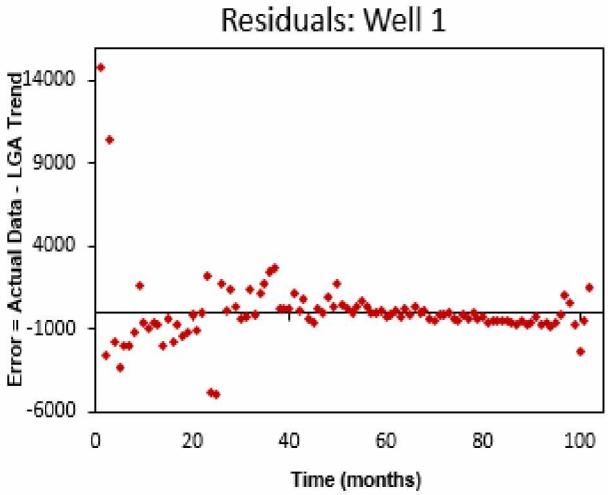

Figure 2.2: Stationary residual dataset obtained for Well 1 by subtracting LGA production profile from actual production data

2.3.1 Modeling the Residual [Autoregressive (AR) Model]

Consider a time series $X(t)$, comprised of measurements x_t observed at time t. If x_t depends on its previous values, p, it is said to be an AR(p) model and can be represented as a linear function of previous values of the series in form of **Eq. 4** (Al-Fattah 2005).

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t \tag{4}$$

ϕ_1, ϕ_2,.. ϕ_p represent autoregressive model parameters at each time lag, and ε_t represents the white noise term. Accordingly, for the purpose of illustration, AR (1) and AR (2) processes can be modeled using **Eqs. 5a** and **5b,** respectively.

$$x_t = \phi_1 x_{t-1} + \varepsilon_t \tag{5a}$$

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t \tag{5b}$$

2.3.2 Modeling the Residual [Moving Average (MA) Model]

If x_t depends on its q previous random error terms (ε_t), it is said to be an MA (q) process and can be represented by **Eq. 6 (**Al-Fattah 2005).

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \tag{6}$$

$\theta_1, \theta_2, \ldots \theta_q$ represent the moving average model parameters at each time lag. Likewise, MA (1) and MA (2) processes can be modeled using **Eqs. 7a and 7b,** respectively.

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} \tag{7a}$$

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} \tag{7b}$$

2.3.3 Modeling the Residual [Autoregressive Moving Average (ARMA) Model]

While the AR and MA models can be used for many data sets, they are not adequate for some data, and a more general model is needed. If x_t depends on its previous values, p, and previous random error terms, q, it is said to be an ARMA (p, q) process and can be represented by **Eq. 8 (**Al-Fattah 2005). All the parameters are as described previously.

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \tag{8}$$

CHAPTER 3 LITERATURE REVIEW

3.1 Analytical Decline Curve Models for Uncertainty Quantification

Several analytical decline curve techniques have been designed for forecasting production and determining uncertainty in shale gas plays (Valko and Lee, 2010; Anderson et al., 2010). Valko and Lee (2010) employed the Stretched Exponential Production Decline (SEPD) model to develop production forecasts for a group of wells. However, forecasts from the analysis of different production periods were combined to quantify uncertainty for a group of wells. Thus, wells with more production history had more influence on the results. Therefore, it is unclear whether the uncertainty quantified in this model is due to the prediction time, the developed model, or an inherent characteristic of the production data.

Anderson and others (2010) developed a decline curve analysis (DCA) technique based on a linear-flow model introduced by Wattenbarger and others (1998). To quantify uncertainty, decline curves generated for different assumed matrix permeability ranges were analyzed probabilistically. However, the technique was only tested on three wells, which is not a substantial sample size to check the method's proficiency.

3.2 Bootstrap and Modified Bootstrap Method (MBM) for Uncertainty Quantification

Some analysts have employed bootstrapping techniques to quantify uncertainty in reserve estimates made by deterministic models. Jochen and Spivey (1996) first applied bootstrapping techniques to generate P_{90}, P_{50}, and P_{10} confidence levels or an 80% confidence interval (CI) for reserve estimates generated by Arps' DCA. They generated synthetic data sets from original production data using conventional bootstrap methods. The work was flawed in its assumption that production data is independent and identically distributed, since production data is time dependent or is a time series.

Cheng and others (2010) introduced Modified Bootstrap Methodology (MBM), which assumed production data to be time dependent, and thus generated synthetic production datasets in a more intuitive manner. Both techniques were tested on a sample size of 100 wells and compared (Cheng et al., 2010). The P_{90}-P_{10} range or 80% CI generated from Cheng and others' (2010) MBM covered true production 80% of the time, whereas Jochen and Spivey's (1996) conventional bootstrap

covered true value only 40% of the time. Therefore, the MBM is a more reliable and rigorous method than conventional bootstrap. Although MBM is well calibrated probabilistically, it is a time consuming process for quantifying uncertainty (Gong et al., 2014). Moreover, both techniques modify the original production data in some manner, which should be avoided if possible.

3.3 Bayesian Markov Chain Monte Carlo Method (MCMC) for Uncertainty Quantification

Bayes' theorem is regarded as one of the most useful statistical inferential theories (Gelman et al., 2004). Gong and others (2014) integrated Bayesian statistics of prior parameter distributions and multiple history matched models to quantify uncertainty in reserve estimates. This modeling technique had three crucial components: the prior distribution for Arps' DCA parameters, which was assumed to be both uniform and non-uniform; the likelihood function, which was determined based on the assumption that error between the logarithmic functions of actual production and estimated production from Arps' DCA will follow a normal distribution; and the Markov Chain Monte Carlo (MCMC) methodology, which is a random sampling algorithm used to optimize distribution for DCA parameters from a proposal distribution. The uncertainty was quantified by integrating the above functions and MCMC algorithm with actual production data available to generate P_{90}, P_{50}, and P_{10} confidence levels for the production forecasts. Although a reliable technique, software packages for implementing the MCMC technique in this context are not readily available in the public domain.

3.4 Time Series Analysis Techniques for Uncertainty Quantification

Time-series analysis techniques have been widely recognized as a stochastic tool for forecasting in fields like economics (Torkowei et al., 2012), epidemiology, medicine, and engineering (Shumway and Stoffer, 2011). In the petroleum industry, Al-Fattah (2005) and Olominu and Sulaimon (2014) have tested the Box and Jenkins Auto Regressive Integrated Moving Average (ARIMA) method for forecasting U.S natural gas production. They employ a purely stochastic ARIMA methodology that uses a differencing technique to make their production dataset stationary for further analysis, which differs from the presented methodology where the trend is determined, and subtracted from the actual dataset to ensure the dataset is at least approximately

stationary. Although the results from their work were in agreement with actual data, their applicability was not tested on a substantial sample size.

CHAPTER 4 METHODOLOGY

4.1 Workflow Overview

The procedure/workflow followed in this study is divided into two stages. The first stage involves estimating LGA model parameters for the sample production dataset using two different non-linear regression schemes: ordinary least square regression (OLS) and weighted least square regression (WLS). The input for the second stage of the study is the two stationary residual datasets generated by subtracting LGA production estimates from actual production in the first stage. These residual datasets are analyzed and modeled using the time series analysis methodology developed by Box and Jenkins (1970) in order to generate the 80% CI for the production forecasts. The overall procedure is shown in **Fig. 4.1** and is described as follows:

- Stage 1: Trend determination and de-trending
    - Decide what fraction of actual production history is to be assumed known. The remainder of the dataset is treated as unknown and used later for model validation.
    - Use OLS and WLS regression schemes on known production history to develop two fitting LGA models.
    - Generate two stationary residual datasets by subtracting predicted production profile from the two developed LGA models.

- Stage 2: Time Series Analysis
    a. Model identification
        - An iterative process where the type and order of ARMA (p ,q) model to be used for generating 80% CI is determined.
    b. Model estimation
        - Coefficients of the identified ARMA model are estimated.
    c. Forecasting with uncertainty bounds
        - The developed ARMA and LGA models are used to generate P_{90}, P_{50}, and P_{10} confidence bounds for production forecasts.

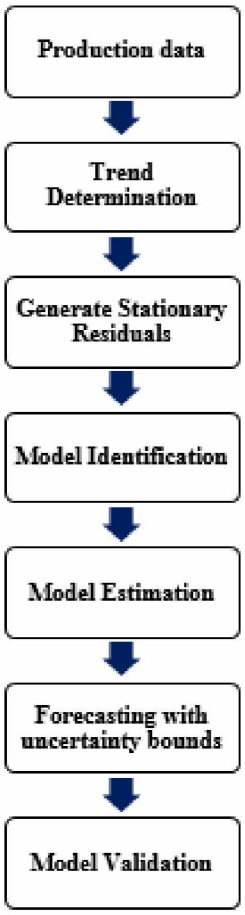

Figure 4.1: Flow diagram for presented Time Series Methodology

4.2 Stage 1: Trend Determination and De-trending

Based on the apparent trend present in production datasets from the Barnett Shale (**Figs. 4.2** to **4.7**), they are obviously non-stationary time series, that is, they contain a trend. Each dataset has 100 months of data, of which only 30 months is assumed known; the rest is used for validating the forecasts. The LGA model was used to define the trend in the data approximately (Eq. 2). The parameters in Eq. 2 were estimated using two non-linear regression schemes. The non-linear regression schemes are:

Ordinary least square (OLS) regression: The model parameters were estimated by minimizing (after several iterations) the square of the difference between the actual data point and the value (response) predicted by the approximated LGA model, conditional to production data.

Mathematically, for n number of data points considered, OLS provides an optimum model when the sum of the residuals, S, is a minimum:

$$S = \sum_{i=1}^{n} w_i (y_i - y_i^{"})^2 \tag{9}$$

Where w_i = weighting factor ($w = 1$ for OLS), y_i = observed response (actual production), and $y_i^{"}$ = predicted response (from the assumed model).

OLS regression assumes equal weight for all data points while estimating the optimal LGA trend model. In tight gas reservoirs, the transient flow period lasts longer than it does in conventional reservoirs. Arps' DCA fails to match production data (results in b>1) from tight gas wells due to the prolonged durations of transient flow (Fetkovich et al., 1987). Intuitively, it would be better to incorporate a second regression scheme where the latest production data, which will have lower values than the transient data, will have more influence on parameter estimation.

Weighted least square (WLS) regression: The regression curves generated by the OLS method may have a tendency to favor higher value data points over lower ones. In WLS, we can apply certain weights in **Eq. 9** to generate regression curves that favor lower value data points that correspond to the latest production data available for analysis. Weighting factors that can place more emphasis on smaller value data points include: $w_i = 1/y_i$, which nearly cancels out weighting of higher amounts and $w_i = 1/y_i^2$, which causes over proportional weighting of smaller amounts.

For the purpose of this study, the weighting factor used for the WLS regression was $w_i = 1/y_i$. $w_i = 1/y_i^2$ was not used because of its tendency to disproportionately weight the more recent data points during regression. Both the WLS and the OLS regression techniques were tested on several production datasets from the Barnett Shale. The wells evaluated each had 100 months of production history. In each analysis, 30 months of data was utilized for developing the LGA model, while the remaining production history was used for verification. It was observed that in some cases, the WLS regression curves matched the production history better than the OLS did (**Figs. 4.2** and **4.3**). In other cases, there was no significant difference (**Figs. 4.4** and **4.5**), while in the

rest, the OLS regression scheme performed better (**Figs. 4.6** and **4.7**). Results are summarized in Table **4.1**.

On the basis of these observations, results from both regression techniques were incorporated into the proposed methodology.

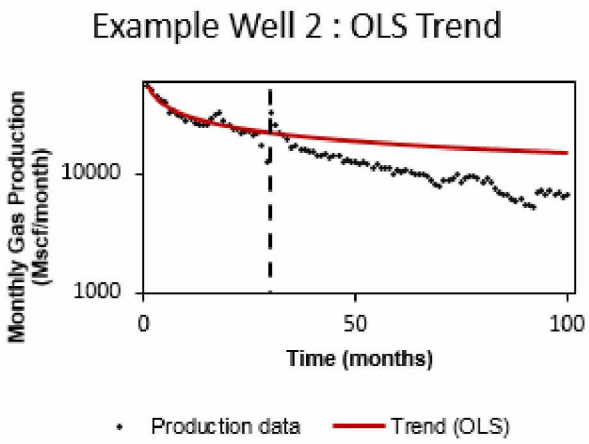

Figure 4.2: LGA trend generated using OLS regression and 30 months of production data for example well 2 (semi-log) (Source: DI Desktop)

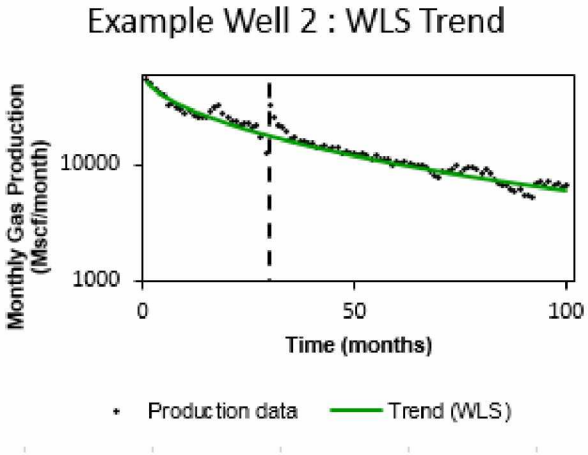

Figure 4.3: LGA trend generated using WLS regression and 30 months of production data for example well 2 (semi-log) (Source: DI Desktop)

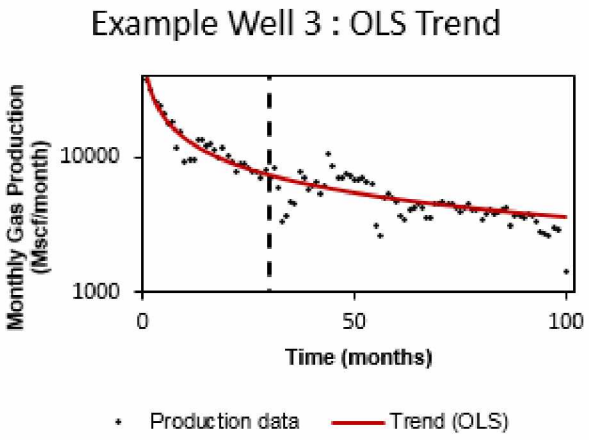

Figure 4.4: LGA trend generated using OLS regression and 30 months of production data for example well 3 (semi-log) (Source: DI Desktop)

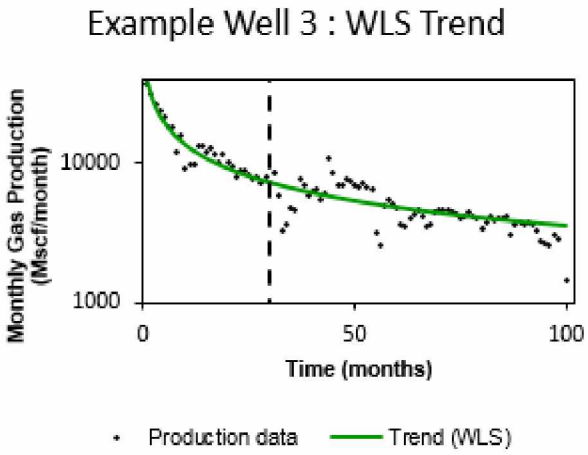

Figure 4.5: LGA trend generated using WLS regression and 30 months of production data for example well 3 (semi-log) (Source: DI Desktop)

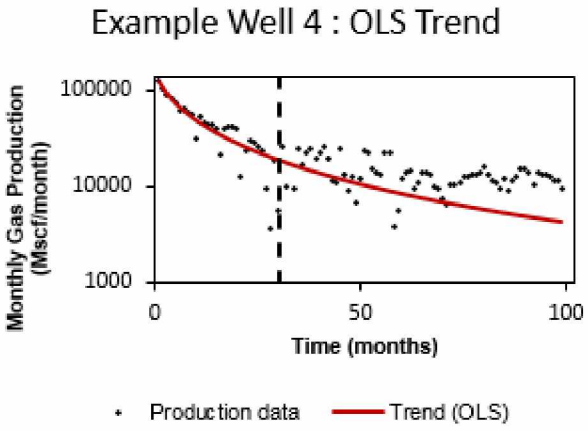

Figure 4.6: LGA trend generated using OLS regression and 30 months of production data for example well 4 (semi-log) (Source: DI Desktop)

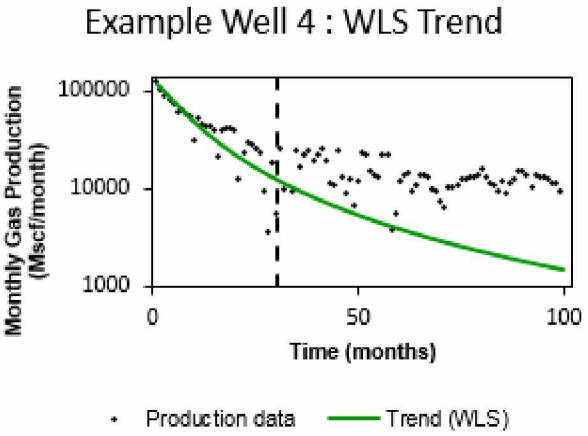

Figure 4.7: LGA trend generated using WLS regression and 30 months of production data for example well 4 (semi-log) (Source: DI Desktop)

Table 4.1: LGA parameters estimated using WLS and OLS regression scheme (30 months known production)

Example Wells	OLS Parameters	WLS Parameters
Well 2	K = 8.1 bcf n = 0.72 a = 83.5	K = 2.92 bcf n = 0.87 a = 41.1
Well 3	K = 0.70 bcf n = 0.94 a = 36	K = 0.72 bcf n = 0.93 a = 36.16
Well 4	K = 2.95 bcf n = 0.81 a = 13	K = 1.87 bcf n = 1.13 a = 15.22

The final step in Stage 1 involves subtracting the WLS and OLS trends from the known production history in order to generate two stationary residual series for each production dataset. Each residual dataset will generate a Confidence Interval (CI), which will be incorporated into our final step. The residual data sets generated for two LGA trends in example well 2 are shown in **Figs. 4.8** and **4.9**.

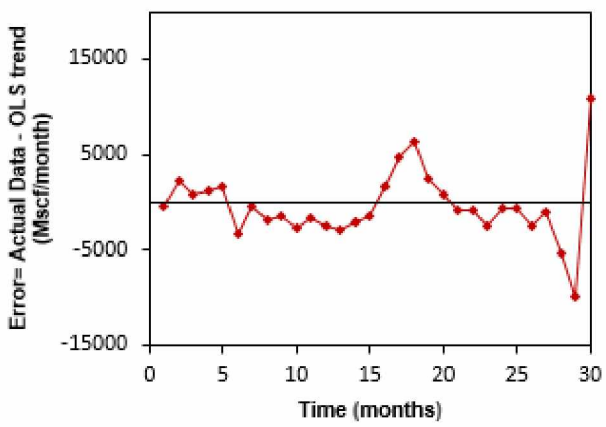

Figure 4.8: Approximately stationary residuals for t=0 to 30 months generated using OLS LGA trend for example well 2

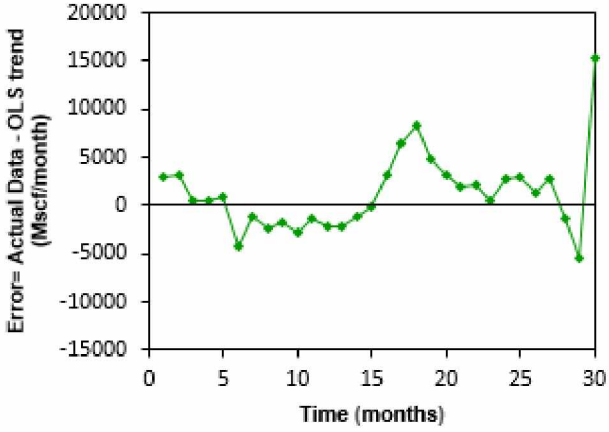

Figure 4.9: Approximately stationary residuals for t=0 to 30 months generated using WLS LGA trend for example well 2

4.3 Stage 2: Time Series Analysis

4.3.1 Model Identification

The first step in the development of a time series model is to identify the order of the ARMA (p , q) model that best fits the data structure of the stationary residual dataset. The autocorrelation function (ACF) and the partial autocorrelation function (PACF) were used to determine the order of p and q . The ACF measures the correlation of a time series with previous and future values. The ACF measures the linear dependency of a variable with itself at two points in time. For a stationary residual time series, autocorrelation between any two observations depends only on the time lag h between them. The sample autocorrelation $\rho(h)$ for a time series $X(t)$ of n>1

observations is given by **Eq. 10** (Shumway and Stoffer, 2011). The numerator represents the auto-covariance at lag h, while the denominator is the auto-covariance at lag 0.

$$\rho(h) = \frac{\sum_{t=1}^{n-|h|}(x_{t+|h|} - \bar{x})(x_t - \bar{x})}{\sum_{t=1}^{n-|h|}(x_t - \bar{x})^2} \tag{10}$$

Where $\bar{x}$ is the mean of the stationary process.

The PACF is an extension of the ACF and is used to examine serial dependencies for individual lags. PACF is essentially the autocorrelation of a signal with itself at different points in time, with linear dependency with that signal at shorter lags removed. PACF is also represented as a function of lags between points in time. Mathematically, the partial correlation between x_t and x_{t+h} is the autocorrelation between x_t and x_{t+h} without the contribution of $x_{t+1}, x_{t+2}, \ldots, x_{t+h-1}$. **Appendix A** details the calculation of ACF and PACF.

Both the ACF and the PACF were calculated numerically using the proper functions in the programming language R. Identifying the best possible ARMA model based on ACF and PACF plots is not an exact science, but rather a process of making an informed choice based on available information. **Table 4.2** summarizes the expected behavior of ACF and PACF plots for the different models.

Table 4.2: ACF and PACF behavior for different models conditional to time series data

Conditional Model	ACF	PACF
AR (p)	Tails off gradually	Cuts off after p lags
MA (q)	Cuts off after q lags	Tails off gradually
ARMA (p , q)	Tails off gradually	Tails off gradually

Figs. 4.10 and **4.11** and **Figs. 4.12** and **4.13** show the identical ACF and PACF plots generated for the OLS and WLS residuals datasets, respectively, for example well 2. This may be due to the absence of any significant contrast in residual datasets (t=0 to 30 months) for both regression schemes. Each bar represents the value of either ACF or PACF at a given lag, and the blue line represents the 95% confidence interval for each function. The bar tails off gradually in the ACF plot and cuts off after lag one in the PACF plot. This suggests that an AR (1) model will match the residual data from both OLS and WLS regression successfully. Several observations were made from the analysis for finding the optimum ARMA model for residual forecasting. First, the ACF and PACF plots for the WLS and OLS residual time series were approximately the same. Second, the behavior of almost all ACF and PACF plots for our sample wells matched the behavior for an AR (1) model. Based on these observations, the AR (1) model was chosen for residual forecasting and CI generation.

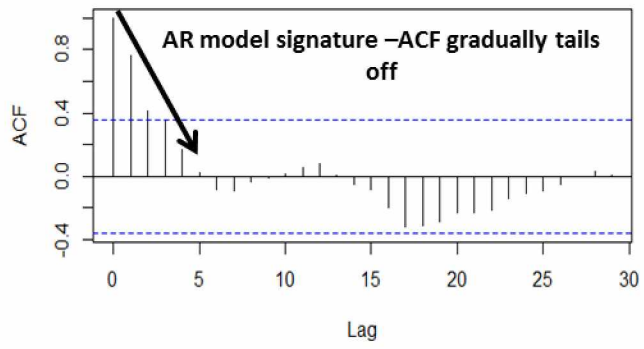

Figure 4.10: ACF plot for residual datasets OLS for example well 2

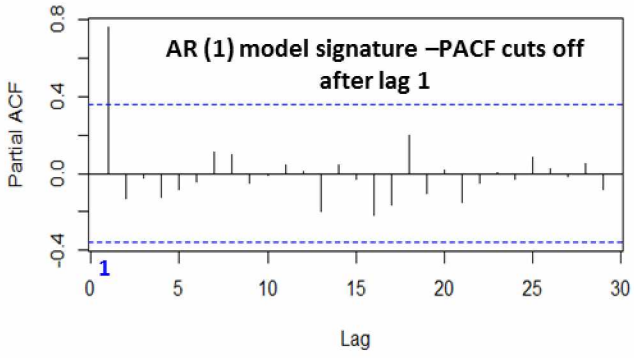

Figure 4.11: PACF plot for residual datasets OLS for example well 2

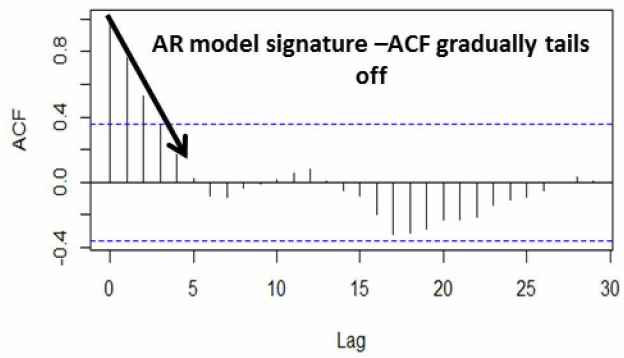

Figure 4.12: ACF plot for WLS residual datasets for example well 2

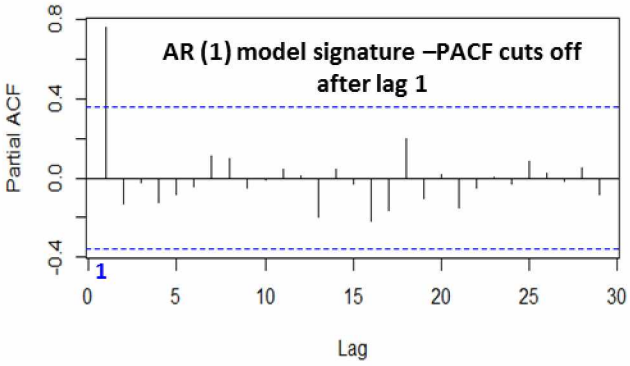

Figure 4.13: PACF plot for WLS residual datasets for example well 2

4.3.2 Model Estimation

In order to estimate the parameters of the AR (1) model, an algorithm was written using the statistical software R. **Eq. 11a** and **Eq. 11b** are the estimated AR (1) models for the OLS and WLS residuals, for example, well 2, in the form of Eq. 5a. Both the models have the same coefficient and mean; the only difference is the variance of the residuals, which is 4201 for OLS and 2500 for WLS. The constant in the equations represents the mean of the sample residual data set.

$$x_t = 0.75x_{t-1} + \varepsilon_t,\ \varepsilon_t \approx (0,4201) \tag{11a}$$

$$x_t = 0.75x_{t-1} + \varepsilon_t,\ \varepsilon_t \approx (0,2500) \tag{11b}$$

4.3.3 Forecasting with Uncertainty Bounds

To clarify the notation of P_{10}-P_{50}-P_{90} in this work, the author assumes that there is a 10%, 50%, and 90% probability that actual reserves are greater than the estimated P_{10}, P_{50}, and P_{90}, respectively. The final step of the process combines the trend forecast made with the two LGA models and the AR (1) residual model. The P_{10}-P_{50}-P_{90} range is generated for both the residual datasets using **Eqs. 11a** and **11b** and incorporated into their respective LGA trend forecasts. **Eqs.**

12a and **12c** represent the generalized form of 80% CI generated by our method as a combination of the trend (LGA) and an error model (AR (1)). These 80% CIs generated for both OLS and WLS regression trends were analyzed and the minimum P_{90} and the maximum P_{10} of the two CIs were chosen as our final P_{90} and P_{10}, respectively. The overall P_{50} is determined by averaging the two P_{50} estimates from our analysis:

$$P_{50}(h) = T(t+h) + R(t+h) \tag{12a}$$

$$P_{10}(h) = T(t+h) + R_{High}(t+h) \tag{12b}$$

$$P_{90}(h) = T(t+h) + R_{Low}(t+h) \tag{12c}$$

Where: h = time step for forecast period (h=1 to 70 months for well 2); t = time step for analysis period (t=30 months for well 2); T = LGA trend component (could be OLS or WLS regression equation); R = residual prediction from estimated AR(1) model; R_{High} = P_{10} residual prediction from estimated AR(1) model; and R_{Low} = P_{90} residual prediction from estimated AR(1) model.

Figs. 4.14 and **4.15** represent the overall P_{10}-P_{50}-P_{90} CI for two different wells with respective forecast periods of 70 months (30 months of data used) and 60 months (40 months of data used). The P_{50} matches the true production better for well B than for well A. However, the true production is within the predicted confidence bounds in both cases.

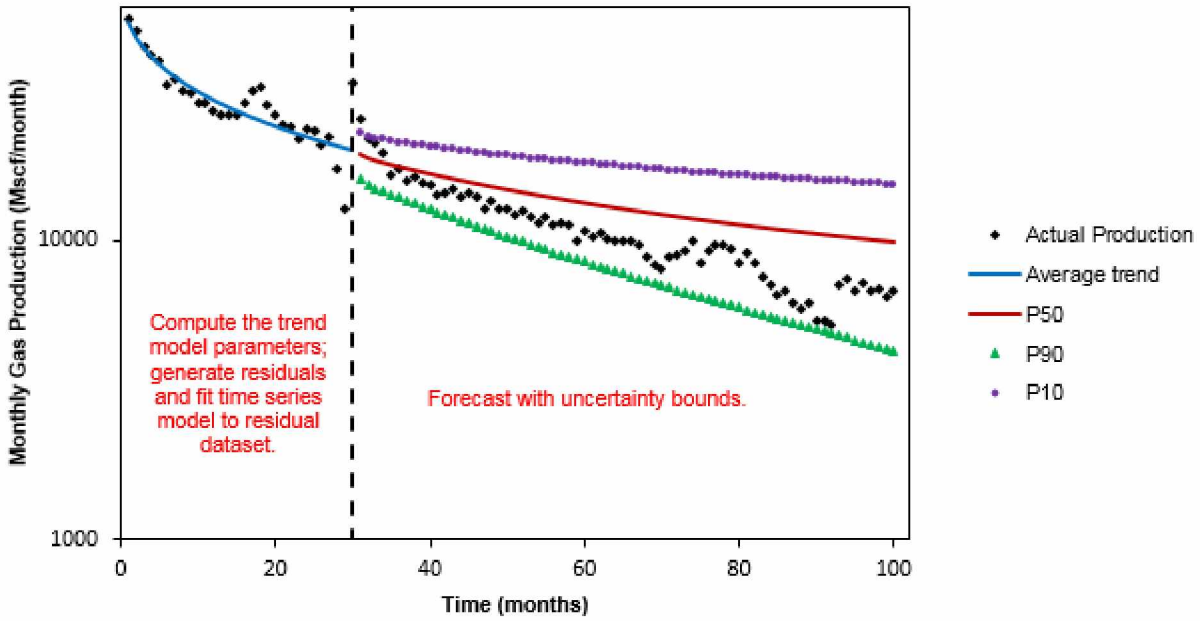

Figure 4.14: Overall P10-P50-P90 confidence levels generated by our methodology for t=70 months for Well A (semi-log) (Source: DI Desktop)

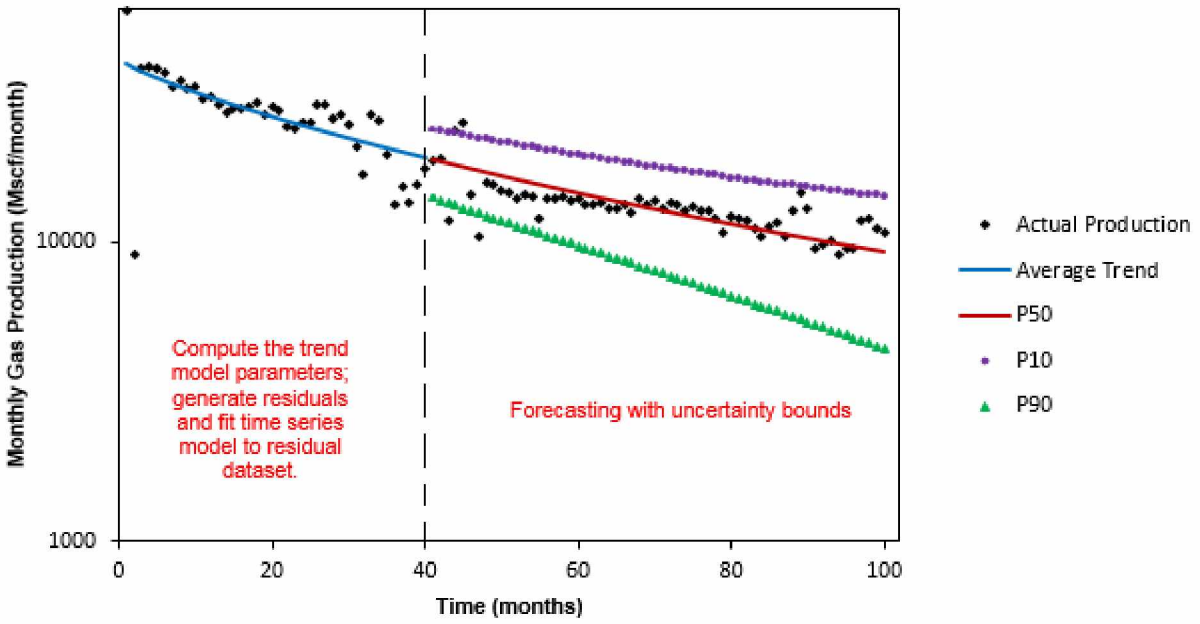

Figure 4.15: Overall P10-P50-P90 confidence levels generated by our methodology for t=60 months for Well B (semi-log) (Source: DI Desktop)

CHAPTER 5 RESULTS AND DISCUSSIONS

5.1 Case Studies

In this section, the methodology is tested on two groups of datasets, described in detail below. The results are examined to check whether the CIs generated are reliable, and determine that the model is well calibrated probabilistically.

5.1.1 Case Study 1: Eight Barnett Shale Counties

A step-by-step workflow has been introduced to generate 80% CI using Time Series analysis and LGA on production data. This methodology is tested on a group of wells from the Barnett Shale, as it is one of the oldest shale gas producing provinces in the country, wherein wells are developed using the latest horizontal drilling and multi-stage hydraulic fracturing techniques. The sample dataset is classified into three groups based on the amount of production history available (**Table 5.1**). Group 1 consists of wells that commenced production in June 2006 (production history = 102 months) from the Denton, Johnson, and Tarrant counties; Group 2 consists of wells that commenced production in June 2007 (production history = 90 months) from the Parker and Wise counties; and Group 3 consists of wells that commenced production in June 2008 (production history = 78 months) from the Erath, Hill, and Hood counties. **Fig. 5.1** showcases monthly production data for each of the 13 wells from Denton County that commenced production in June 2006. The median production for each month (**Fig. 5.2**) is calculated because the data is less chaotic and relatively easier to analyze. Such median datasets were generated and analyzed by the methodology described previously for all the counties in our sample. A fraction of the production data is used for trend and residual model parameter estimation, as shown in Table 3.

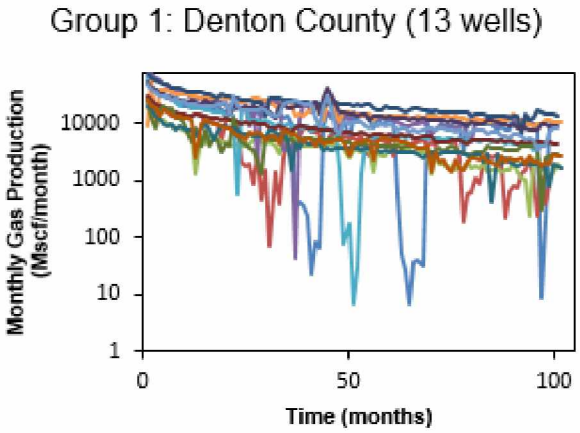

Figure 5.1: Monthly production data from 13 wells in Denton county on semi-log plot (Production start: June 2006) (Source: DI Desktop)

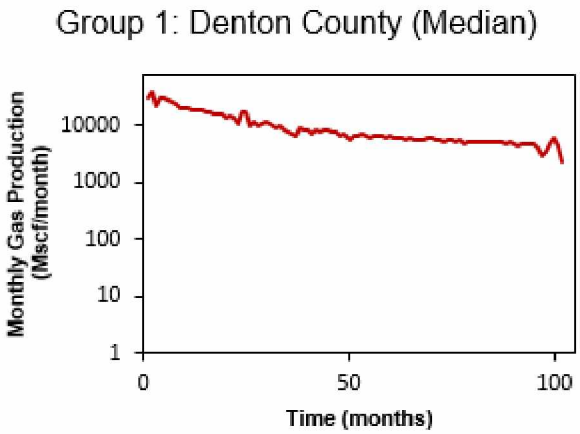

Figure 5.2: Median of production data from 13 wells of Denton county on semi-log plot (Production start: June 2006) (Source: DI Desktop)

For each group of wells only a portion of production history was assumed as known (Table 5.1). Then, forecast is made for the remaining production from the last date of analyzed production to the last date of actual production. An analysis is performed where the estimated P_{10}-P_{50}-P_{90} cumulative production at the end of the production period was compared with the actual cumulative production (**Table 5.2**). In each case, the P_{10}-P_{90} range brackets the actual monthly production successfully for medium to long range application, as shown in **Figs. 5.3** through **5.10**. The legends are detailed in Fig. 5.3 and are applicable to Figs. 5.4 to 5.10. The CI generated varies in width for each county based on the variance of production data utilized and the contrast in the trends generated by WLS and OLS regressions.

Table 5.1: Case Study 1 dataset, with number of wells and assumed known production from each county

Group	County	Number of Wells	Assumed Known Production History	Production history for validation
A	Denton	13	40 months	62 months
	Johnson	21	40 months	62 months
	Tarrant	19	40 months	62 months
B	Parker	25	30 months	60 months
	Wise	13	30 months	60 months
C	Erath	7	20 months	58 months
	Hill	10	20 months	58 months
	Hood	8	20 months	58 months

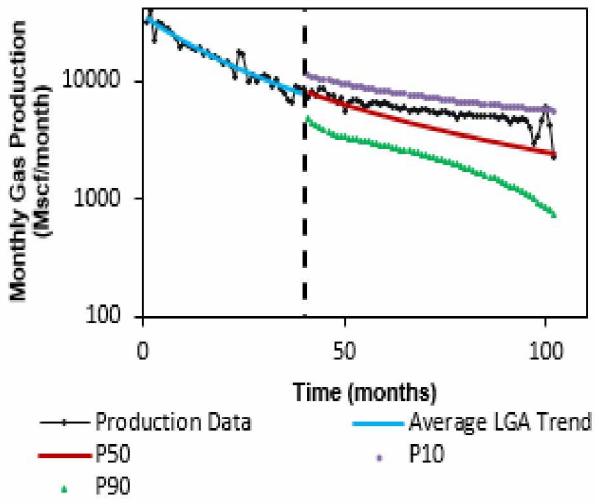

Figure 5.3: 80% CI prediction for Denton County on semi-log plot (Forecast period = 62 months) (Source: DI Desktop)

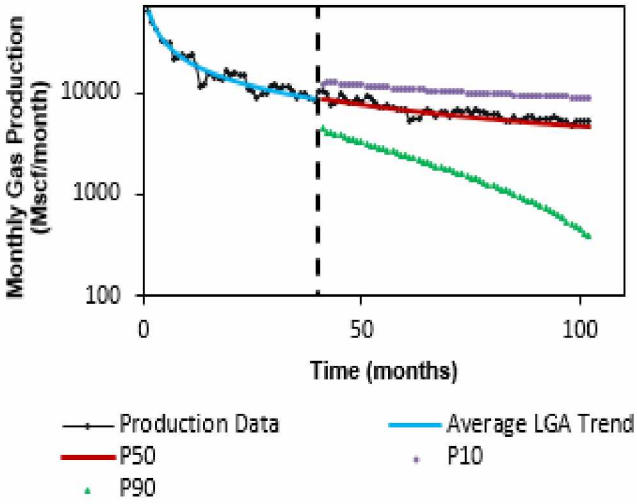

Figure 5.4: 80% CI prediction for Johnson County on semi-log plot (Forecast period = 62 months) (Source: DI Desktop)

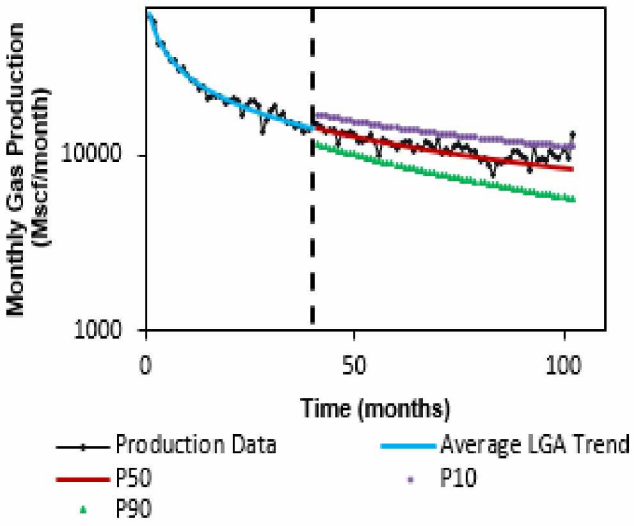

Figure 5.5: 80% CI prediction for Tarrant County on semi-log plot (Forecast period = 62 months) (Source: DI Desktop)

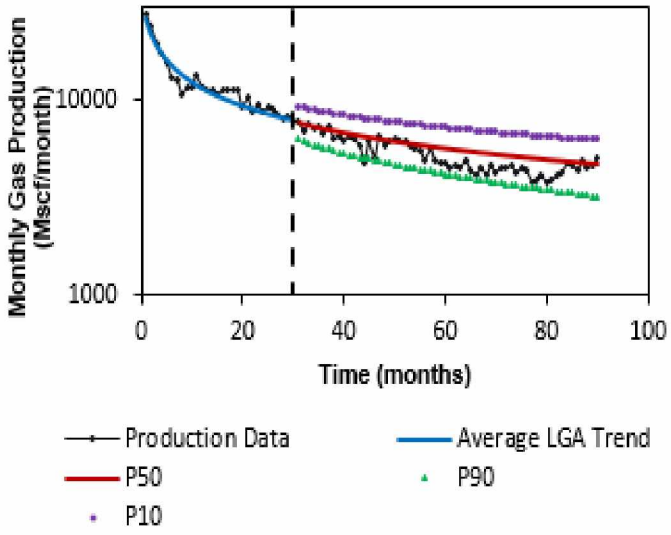

Figure 5.6: 80% CI prediction for Parker County on semi-log plot (Forecast period = 60 months) (Source: DI Desktop)

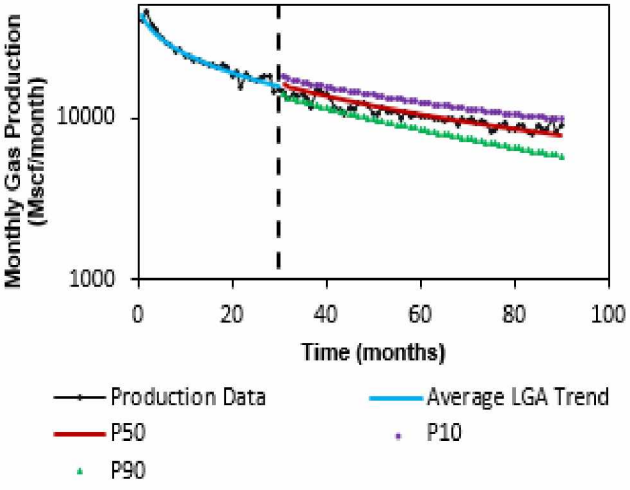

Figure 5.7: 80% CI prediction for Wise County on semi-log plot (Forecast period = 60 months) (Source: DI Desktop)

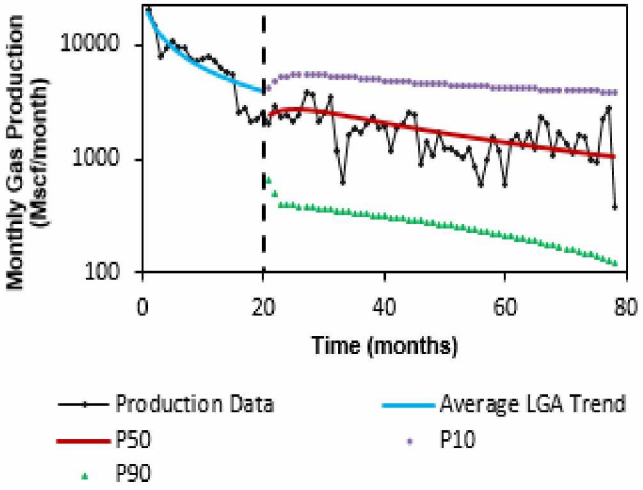

Figure 5.8: 80% CI prediction for Erath County on semi-log plot (Forecast period = 58 months) (Source: DI Desktop)

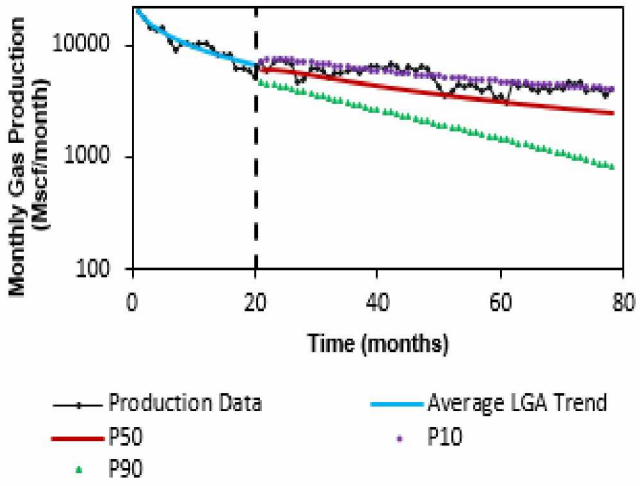

Figure 5.9: 80% CI prediction for Hill County on semi-log plot (Forecast period = 58 months) (Source: DI Desktop)

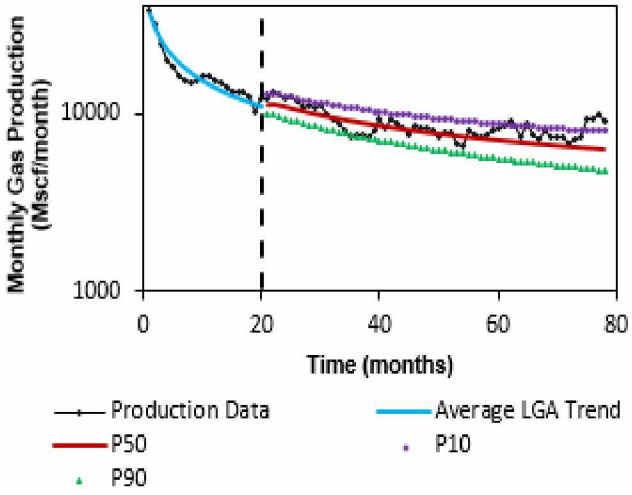

Figure 5.10: 80% CI prediction for Hood County on semi-log plot (Forecast period = 58 months) (Source: DI Desktop)

Table 5.2: Comparison of P10-P50-P90 confidence bound generated by our method for cumulative production data at the end of production period, with actual cumulative production (Case Study 1)

Group A	AR(1) model (Data analyzed=40 months)	Variance of error (ε_t)		Cumulative Production at end of 102 months			
		OLS	WLS	P10 - P90 (Mmcf)	P50 (Mmcf)	Actual (Mmcf)	Relative Error %
Denton	0.22	2620293	2690014	1.13 – 0.74	0.93	1	0.7
Johnson	0.38	3923562	4037358	1.35 – 0.83	1.09	1.11	1.8
Tarrant	0.17	1951850	1939625	1.81 – 1.47	1.64	1.66	1.2
Group B	AR(1) Coefficient (Data analyzed=30 months)	Variance of error (ε_t)		Cumulative Production at end of 90 months			
		OLS	WLS	P10 - P90 (Mmcf)	P50 (Mmcf)	Actual (Mmcf)	Relative Error %
Parker	0.26	572416	572416	0.81 – 0.63	0.72	0.68	5.5
Wise	-0.28	1086617	957703	1.49 – 1.25	1.37	1.35	1.5
Group C	AR(1) Coefficient (Data analyzed=20 months)	Variance of error (ε_t)		Cumulative Production at end of 78 months			
		OLS	WLS	P10 - P90 (Mmcf)	P50 (Mmcf)	Actual (Mmcf)	Relative Error %
Erath	0.77	825507	772525	0.41 – 0.15	0.25	0.24	4.2
Hill	0.67	391300	380096	0.53 – 0.34	0.43	0.5	14
Hood	0.24	607080	607080	0.91 – 0.73	0.82	0.85	3.5

To check for time series model adequacy, It is tested whether ε_t in each case was white noise. This is called AR Model Diagnostics. The value of ε_t was calculated by subtracting the AR(1) model prediction from the actual residuals. White noise is a process with no autocorrelation, hence the ACF of such a process dies after lag 0 and the PACF plot is flat zero. The behavior of the ACF and PACF plots for ε_t of all the sample counties in Case 1 was similar to that of white noise, as shown in **Figs. 5.11** and **5.12**.

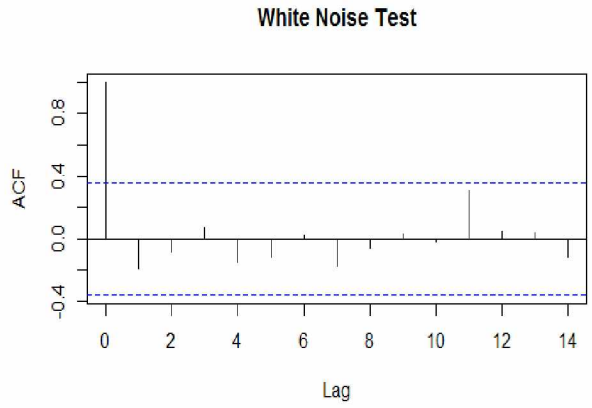

Figure 5.11: ACF plot for white noise testing of ε_t for Denton County

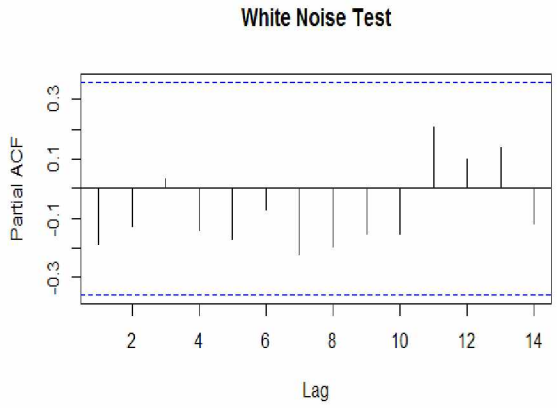

Figure 5.12: PACF plot for white noise testing of ε_t for Denton County

5.1.2 Case Study 2: 100 Barnett Shale Wells

To check the reliability of the 80% CI generated by our method, the presented technique was tested on 100 individual Barnett Shale wells with at least 100 months of production history (**Figure 5.13**). For this work, it would be assumed that the method is well calibrated if the true production value is contained within our CI 80% of the time in a sample of 100 wells. Since this method's performance may vary depending upon the amount of production history utilized, 40, 50 and 60 months of production data was used for the calibration test in each individual well. The results from the analysis will highlight any variation in the shape and accuracy of our generated CI with the amount of production history used. The originally acquired data has not been modified in our

work. This is so that the process can be standardized and can be tested successfully with any similar datasets.

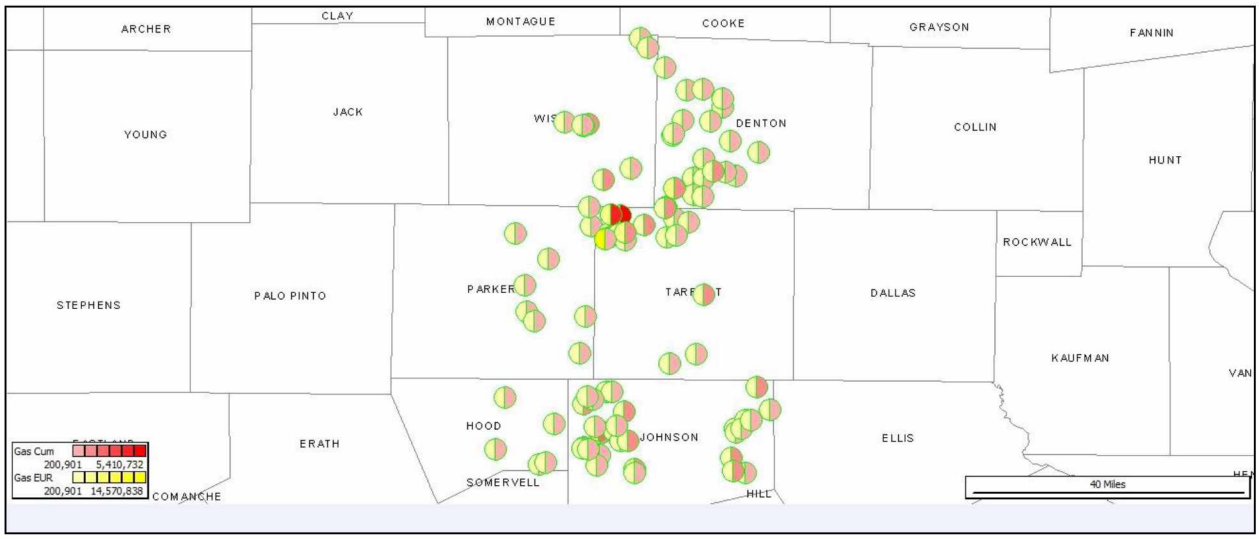

Figure 5.13: Position of 100 sample wells for our calibration test

The method was tested on production data from 100 individual wells in the Barnett Shale for the three described scenarios (40, 50 and 60 months of data utilized) and several observations were made. The coverage probability or coverage rate of a CI is the proportion of the time that the interval contains the true value of interest in a large sample, so the expected coverage rate of P_{10}-P_{90} generated for the production forecast by our method should be 80%. The CIs generated covered true production 84%, 90%, and 92% of the time when 40, 50, and 60 months of data were utilized, respectively (see **Table 5.3**). The realized coverage rate is better than the expected rate of 80%. **Figs. 5.14** through **5.16** are semi-log plots showing variation in the width of CIs for the three scenarios. The CIs get tighter as more production data was used for analysis, suggesting a decrease in related uncertainty with increasing amount of data utilized. The parameters and variance of the realized AR (1) models for these 100 wells are summarized in **Appendix B.**

Table 5.3: Realized coverage rate for generated 80% CI tested on 100 Barnett Shale gas wells

Coverage Rate (80% CI)	Known Production History		
	40 months	50 months	60 months
Expected	80%	80%	80%
Realized	84%	90%	92%

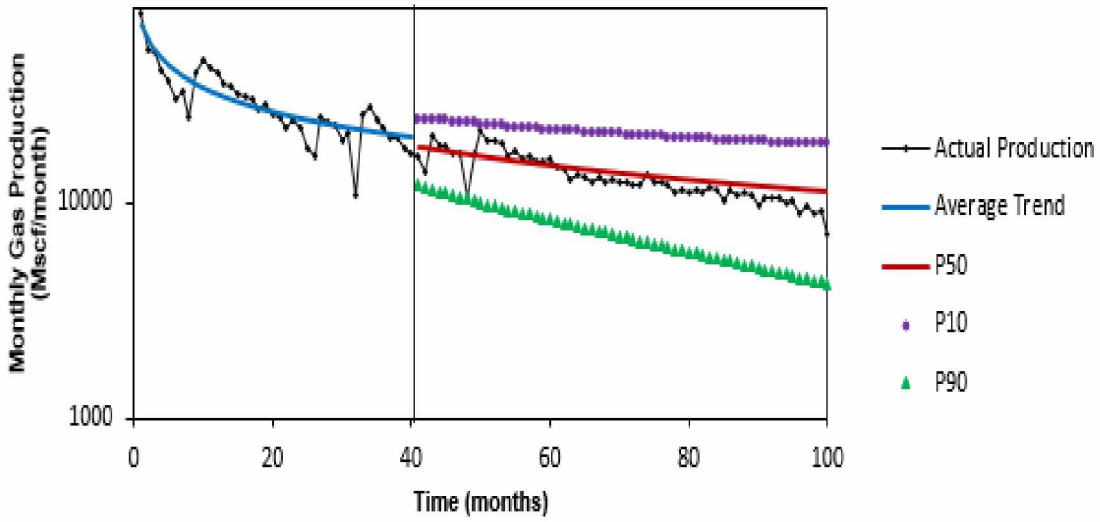

Figure 5.14: 80% CI generated for a sample Barnett Shale well using 40 months production history on a semi-log plot (Source: DI Desktop)

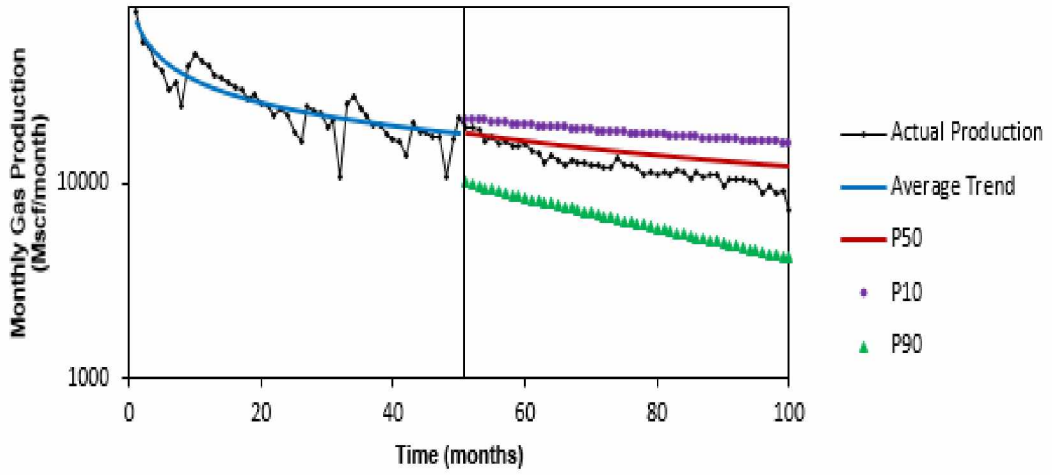

Figure 5.15: 80% CI generated for a sample Barnett Shale well using 50 months production history on a semi-log plot (Source: DI Desktop)

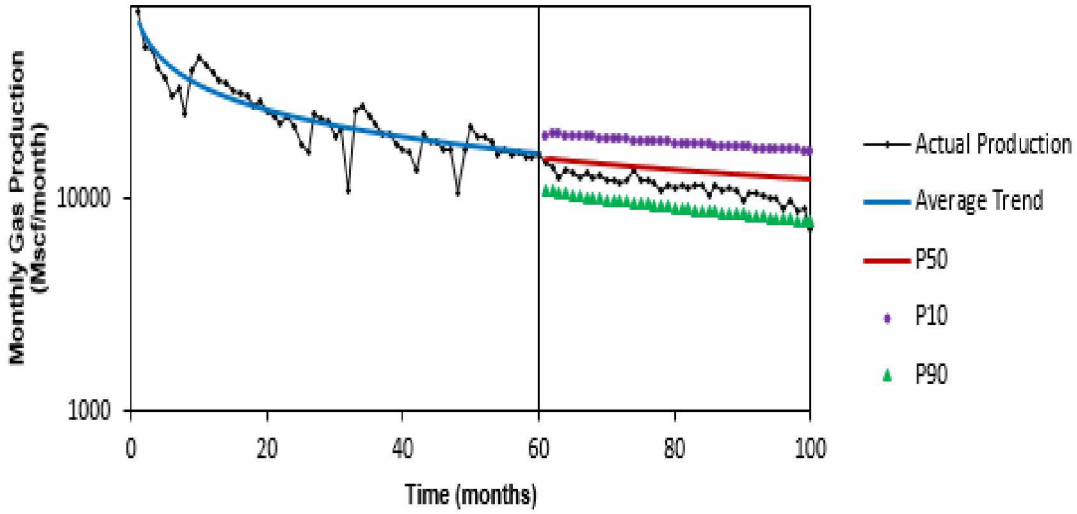

Figure 5.16: 80% CI generated for a sample Barnett Shale well using 60 months production history on a semi-log plot (Source: DI Desktop)

As described earlier, the expected probabilities of observing actual production values greater than P_{10}, P_{50}, and P_{90} are 10%, 50%, and 90%, respectively. For this work, it is considered that the method is well calibrated if the frequencies of "true value > our estimates" are at least 10%, 50%, and 90% for P_{10}, P_{50}, and P_{90}, respectively. **Fig. 5.17** shows results from the analysis of cumulative

production of 100 wells for three scenarios. The conditions for method calibration are approximately satisfied, with optimum results obtained when 60 months of data was utilized.

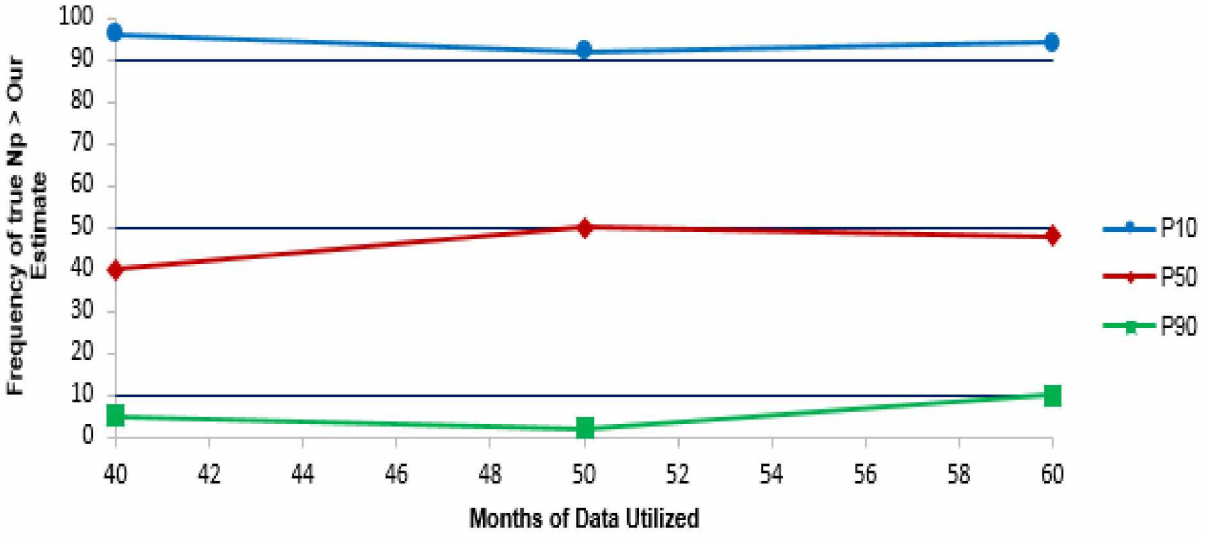

Figure 5.17: Results of 80% CI for 100 wells for testing method calibration

An AR Model Diagnostic needed to be performed for Case 2. The same time series model adequacy test used in Case 1, was performed for results from Case 2. **Figs. 5.18** and **5.19** showcase the ACF and PACF plots of ε_t for one such well. The results suggest that the estimated AR models are in agreement with the residual dataset.

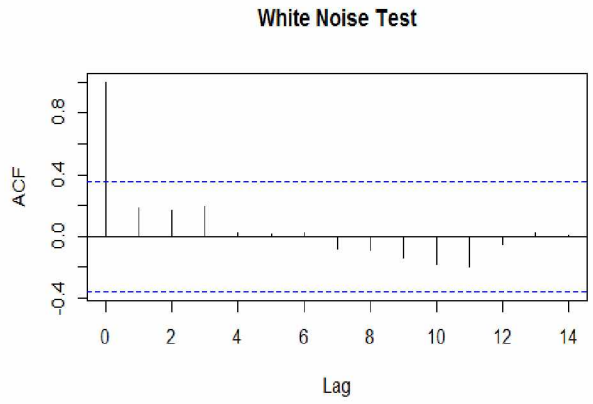

Figure 5.18: ACF plot for white noise testing of ε_t for a sample well

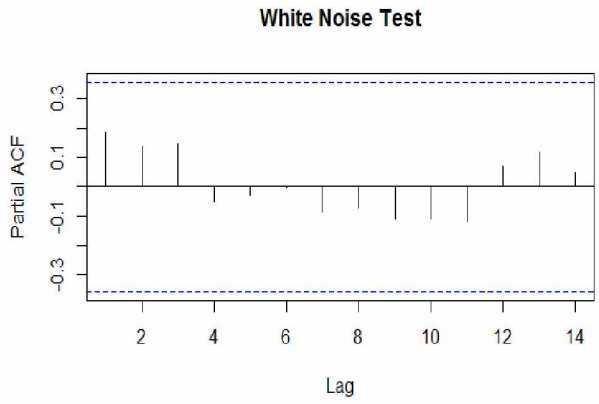

Figure 5.19: PACF plot for white noise testing of ε_t for a sample well

5.2 Comparison with Bayesian MCMC

In this section, the P_{10}-P_{50}-P_{90} bounds generated by presented method are compared with those generated by the Bayesian MCMC method used by Gong and others (2014). As discussed previously, Bayesian MCMC integrates Bayes' theorem with Arps' DCA to generate CIs, whereas MBM uses the modified bootstrapping developed by Cheng and others (2010). **Fig. 5.20** is a semi-log time vs. production rate plot that shows the contrast between P_{10}-P_{50}-P_{90} bounds generated by Bayesian MCMC and MBM. In their work the P_{90}, P_{50}, and P_{10} are representative curves of the 10th, 50th, and 90th percentiles of their forecasts. According to Fig. 15a, both the Bayesian method and MBM bracket actual production successfully inside their P_{90}–P_{10} bounds. The Bayesian P_{90}–P_{10} bounds are narrower than those from MBM for this example. Narrower bounds are more desirable if both techniques quantify uncertainty equally well. **Fig. 5.21** compares the P_{90}–P_{10} bounds generated by Bayesian MCMC with P_{10}-P_{90} bounds generated by our method. It is observed that both the MCMC and our method bracket true production value equally well within the 80% CI for the example well from Gong and others (2014). However, presented works CI is wider for this particular example and tends to bracket slightly more monthly production data points compared to the Bayesian MCMC CI. Therefore, for the showcased well, MCMC is more desirable

than our method due to its tighter bounds, although the 80% CIs by MBM is wider than that of our method.

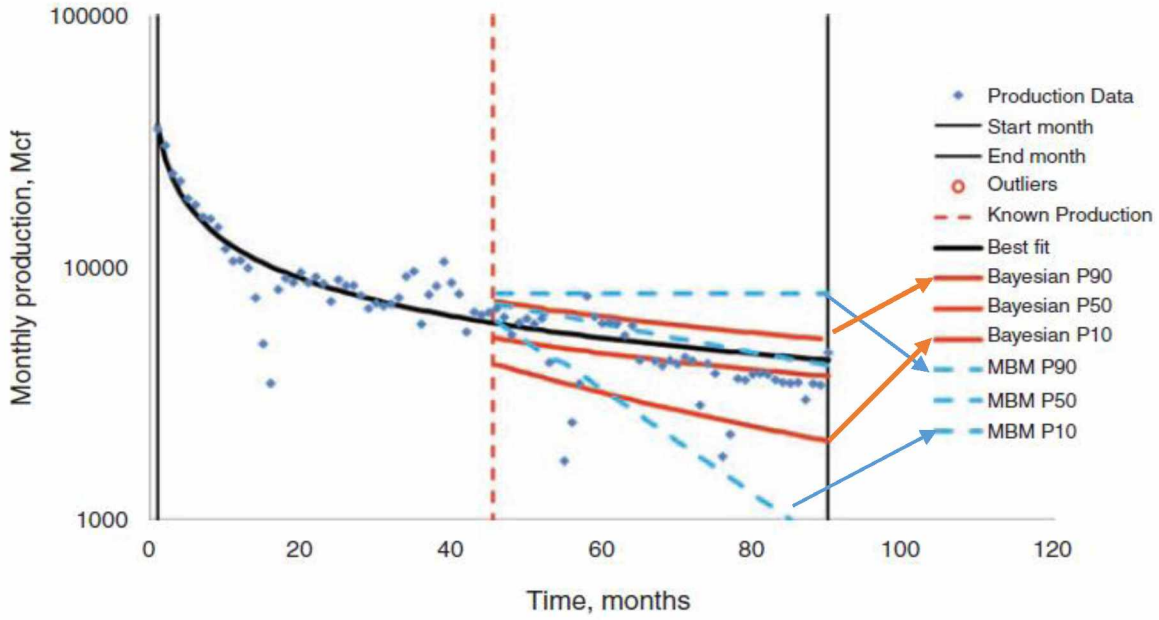

Figure 5.20: MCMC 80% CI generated by Gong and others (2014) on semi-log graph

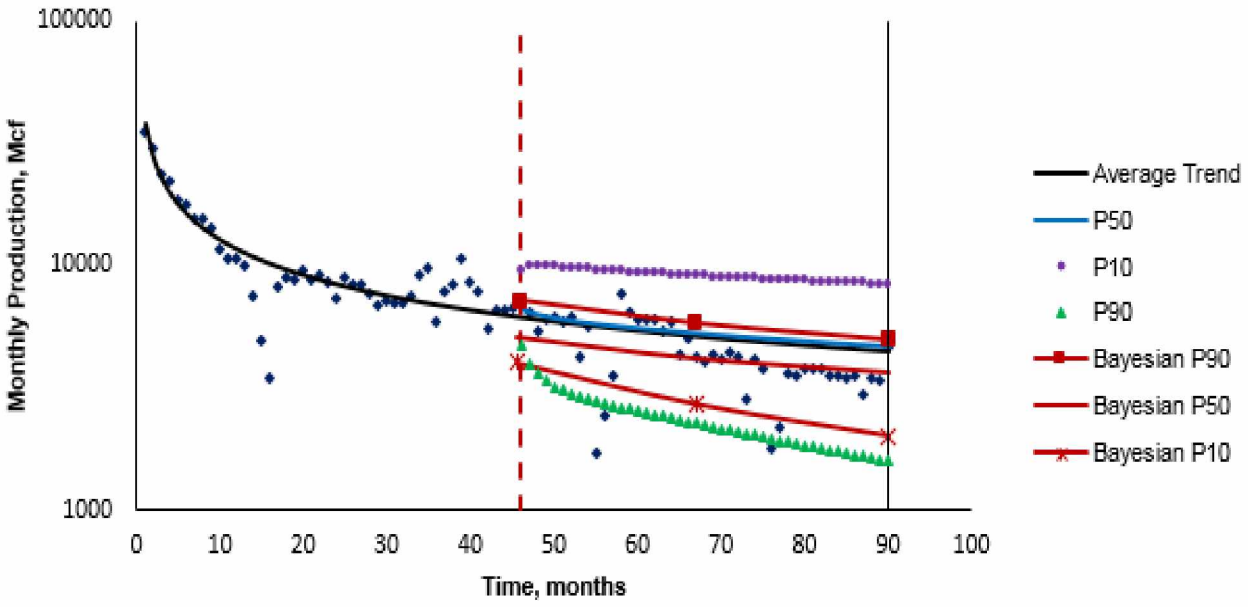

Figure 5.21: Comparison of CI generated by MCMC method and presented method on semi-log graph

5.3 Application of Methodology to Re-fractured Horizontal Well

In some cases, a well is re-fractured to increase productivity, as shown in **Fig. 5.22.** There are is a local peak in the production data, at t= 73 months. As the trend of the dataset changes after the re-fracturing process, The production data is partitioned into two sections and analyzed separately. These sections will be called "data before re-fracturing" and "data after re-fracturing" for convenience. The developed methodology was applied to the first section to generate 80% CI, as showed in **Fig. 5.23**. 30 months of data was utilized to develop the trend and the AR(1) residual model. Since there is a limited amount of data after re-fracturing, only 6 months of data is used to determine the trend. The developed trend was incorporated into the residual model developed for the data before re-fracturing to generate 80% CI, as shown in **Fig. 5.24**. Evidently, the CIs successfully bracketed the true production data before and after re-fracturing for up to 50 months and 20 months, respectively. This technique has utility when a shale gas producer doesn't have sufficient "data after re-fracturing" to generate reliable CI for forecasts. He can utilize the AR(1) model from "data before re-fracturing" combined with the trend from "data after re-fracturing", to generate forecasts with CI even with limited data.

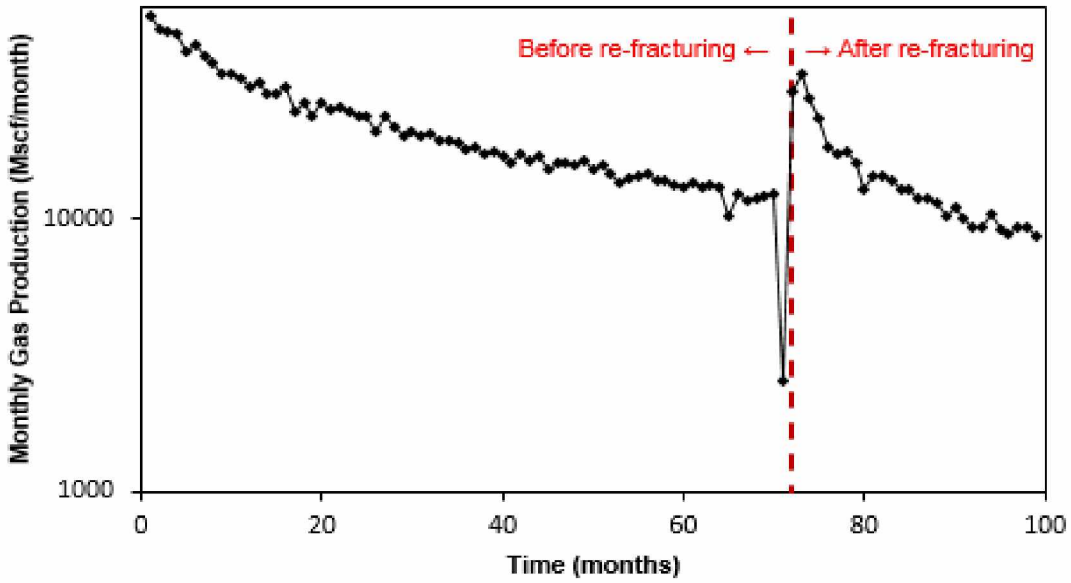

Figure 5.22: Example production data from a re-fractured well (semi-log)

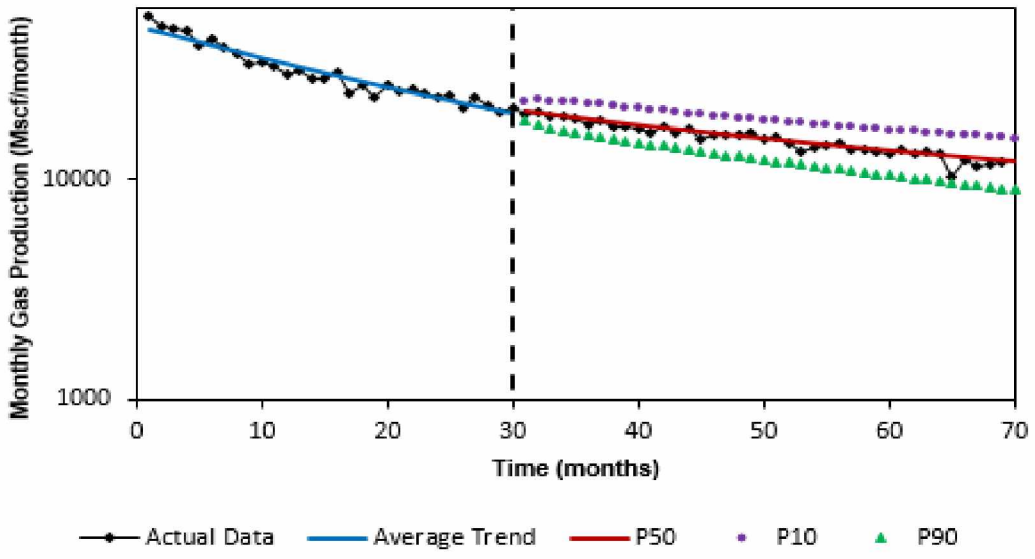

Figure 5.23: 80% CI generated for data before re-fracturing, forecast period= 40 months

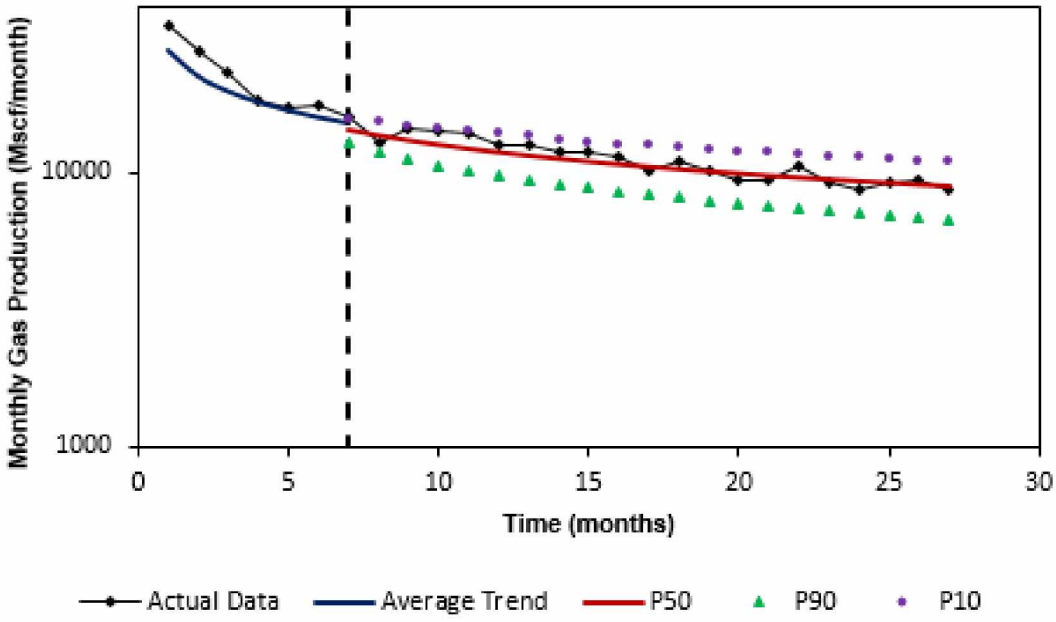

Figure 5.24: 80% CI generated for data after re-fracturing, forecast period= 20 months

CHAPTER 6 CONCLUSIONS AND RECOMMENDATIONS

In the first stage of the presented methodology, LGA parameters were determined for the available production data using ordinary least squares and weighted least squares regressions. The predicted trend from the estimated LGA equation was subtracted from actual production to generate stationary residuals. In the second stage, these stationary residuals were analyzed and modeled using time series analysis. The predicted AR(1) time series model generated for residuals was then combined with the LGA trend model to forecast monthly production with an 80% confidence interval for a sample unconventional gas well. The applicability of this method was tested on two case studies from the Barnett Shale and my analysis of the results led to following conclusions:

- At least two types of regression scheme can be used to fit the production data trend, and their performance may vary on a case-to-case basis. Using available production data, ordinary least squares (OLS) regression provides an optimistic estimate, while weighted least squares (WLS) regression provides a conservative estimate.
- For a test case of groups of wells from eight Barnett Shale counties, time series methodology coupled with LGA quantified uncertainty reliably with as little as 20 months of data.
- The actual coverage rate of P_{10}-P_{50}-P_{90} bounds for a sample dataset of 100 Barnett Shale wells was better than the expected coverage rate of 80%. This also indicates that the method is well calibrated probabilistically.
- The generated confidence intervals got narrower as more data was used for analysis, suggesting a decrease in related uncertainty with more available data.
- The confidence interval generated by our method is wider than those of Bayesian MCMC. This suggests that our method is more permissive, or generates wider bounds (based on results from analysis on one example well).

For history matching applications in this work, it has been demonstrated that this methodology is well calibrated: P_{10}, P_{50}, and P_{90} estimates correspond to realized frequencies of approximately 90%, 50% and 10%, as desired. This suggests that the developed methodology can be applied to production forecasting and uncertainty quantification, especially for unconventional reservoirs. This methodology can be integrated with other deterministic models other than LGA by replacing

the LGA trend equations (Eq. 2) with the appropriate equations. Also, the presented method is computationally inexpensive and easy to implement.

Results from only one kind of weighted regression techniques, along with the ordinary regression technique, has been incorporated in this work. Further research can be conducted to include more types of weighted regression techniques. The applicability of the procedure can be further tested for both conventional oil and gas plays as well.

REFERENCES

Al-Fattah, S.M. 2005. Time Series Modeling for U.S. Natural Gas Forecasting. Presented at the International Petroleum Technology Conference, Doha, Qatar, 21-23 November. IPTC-10592-MS. http://dx.doi.org/10.2523/10592-MS

Anderson, D. M., Nobakht, M., and Moghadam, S., et al. 2010. Analysis of Production Data from Fractured Shale Gas Wells. Presented at SPE Unconventional Gas Conference, Pittsburg, Pennsylvania, 23-25 February. SPE 131787-MS. http://dx.doi.org/10.2118/147588-PA

Arps, J. J. 1945. Analysis of Decline Curves. *Transactions of the AIME* **160** (01): 228 – 247 SPE-945228-G. http://dx.doi.org/10.2118/945228-G

Box, G.E.P. and Jenkins G.M. 1970. Time Series Analysis, Forecasting, and Control. Oakland, CA: Holden-Day

Cheng, Y., Wang, Y., and McVay, D. et al. 2010. Practical Application of a Probabilistic Approach to Estimate Reserves Using Production Decline Data. SPE Econ & Mgmt 2(1): 19-31. SPE 95974-PA. http://dx.doi.org/ 10.2118/95974-PA

Clark, A.J., Lake, L.W. and Patzek, T.W. 2011. Production Forecasting With Logistic Growth Models. Presented at SPE Annual Technical Conference and Exhibition, Denver, Colorado, 30 October-2 November. SPE-144790-MS. http://dx.doi.org/ 10.2118/144790-MS

Duong, A.N. 2011. Rate-Decline Analysis for Fracture Dominated Shale Reservoirs. Presented at Canadian Unconventional Resources and International Petroleum Conference, Calgary, Alberta, 19-21 October. SPE-137748-PA. http://dx.doi.org/10.2118/137748-PA

Drillinginfo (DI) Desktop, version 6.2.1.0. 18 July 2013. http://info.drillinginfo.com/di-desktop-downloads/

Fetkovich, M. J., Vienot, M. E., Bradley, M. D., and Kiesow, U. G. 1987. Decline Curve Analysis Using Type Curves: Case Histories. *SPE Formation Evaluation*: 637-656. SPE 13169. http://dx.doi.org/10.2118/13169-PA.

Gelman, A., Carlin, J. and Stern, H. et al. 2004. Bayesian Data Analysis. Second Edition. Boca Raton, Florida: CRC Press

Gong, X., Gonzalez, R., McVay, D. A., and Hart, J. D. 2014. Bayesian Probabilistic Decline-Curve Analysis Reliably Quantifies Uncertainty in Shale-Well-Production Forecasts. SPE Journal. 19 (06): 1,047 - 1,057. SPE-147588-PA. http://dx.doi.org/10.2118/147588-PA

Ilk, D., Rushing, J. A., Perego, A. D. et al. 2008. Exponential vs. Hyperbolic Decline in Tight Gas Sands: Understanding the Origin and Implications for Reserve Estimates Using Arps Decline Curves. Presented at SPE Annual Technical Conference and Exhibition, Denver, Colorado, 21-24 September. SPE-116731-MS. http://dx.doi.org/10.2118/147588-PA

Jochen, V. A., and Spivey, J. P. 1996. Probabilistic Reserves Estimation Using Decline Curve Analysis with the Bootstrap Method. Presented at SPE Annual Technical Conference and Exhibition, Denver, Colorado, 6-9 October. SPE 36633-MS. http://dx.doi.org/10.2118/36633-MS

Mathuranthan, V. 2014. Yule Walker Estimation and Simulation in Matlab. *gaussianwaves*, 27 May 2014, http://www.gaussianwaves.com/2014/05/yule-walker-estimation/ (accessed 5 August 2015)

Oberkampf, W.L., Deland, S.M., Rutherford, B.M. et al. 2002. Error and Uncertainty in Modeling and Simulation. Reliability Engineering and System Safety 75: 333

Olominu, O., and Sulaimon, A.A. 2014. Application of Time Series Analysis to Predict Reservoir Production Performance. Presented at SPE Nigeria Annual International Conference and Exhibition, Lagos, Nigeria, 05-07 August. SPE-172395-MS. http://dx.doi.org/10.2118/172395-MS

Paryani, M., Ahmadi, M., Awoleke, O. and Hanks, C. 2015. Influence of Residual Function Form on the Performance of Decline Curve Analysis Models for Shale Oil Wells. Manuscript submitted for peer review in the SPE Res Eval & Eng Journal

Shumway, R. and Stoffer, D. 2011. *Time Series Analysis and Its Applications*, third edition. New York: Springer Science and Business Media

Spencer, R.P. and Coulombe, M.J. 1966. Quantification of Hepatic Growth and Regeneration. Growth, Development and Aging 30(3): 277-284

Torkowei, B. V. and Ikiensikimama, S. S. 2012. Understanding Crude Oil Price Variation Using Time Series Analysis. Presented at SPE Nigeria Annual International Conference and Exhibition, Abuja, Nigeria, 06-08 August.SPE-162964-MS. http://dx.doi.org/10.2118/162964-MS

Valko, P. P. and Lee, W. J. 2010. A Better Way To Forecast Production From Unconventional Gas Wells. Presented at SPE Annual Technical Conference and Exhibition, Florence, Italy, 19-22 September. SPE 134231-MS. http://dx.doi.org/10.2118/147588-PA

Verhulst, P.F. 1838. Correspondance mathematique et physique 10: 113-121

Wattenbarger, R.A., El-Banbi, A.H., and Villegas, M.E. et al. 1998. Production Analysis of Linear Flow into Fractured Tight Gas Wells. Presented at SPE Low Permeability Reservoirs Symposium and Exhibition, Denver, Colorado, 5-8 April. SPE 39931. http://dx.doi.org/10.2118/39931-MS

APPENDIX A

ACFs and PACFs are crucial for identifying patterns and subsequently determining the best fitting model for the time series data. Both autocorrelations and partial autocorrelations are computed for sequential lags in series. The first lag will have autocorrelation between x_{t-1} and x_t, the second lag between x_{t-2} and x_t, and so on, as ACFs and PACFs function across all lags. The formula used for calculating the sample ACF for a given stationary time series is Eq. 10.

In this section, the PACF is explained in detail. Before that, let us define a casual AR(1) process given by:

$$X_t = \phi X_{t-1} + Z_t \tag{A-1}$$

where, ϕ is the AR coefficient and Z_t is the white noise process. Then the auto covariance function (ACVF) of $\{X_t\}$ at lag h is:

$$\gamma_X(h) = \text{cov}(X_{t+h}, X_t) = E[(X_{t+h} - \mu_X)(X_t - \mu_X)] \tag{A-2}$$

where, γ_X is the ACVF of time series, μ_X is the mean of the time series, E is the expected value operator, and cov represents the function covariance. The ACF of $\{X_t\}$ at lag h is:

$$\rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)} = cor(X_{t+h}, X_t) \tag{A-3}$$

Now for calculating ACVF at lag 2:

$$\gamma(2) = \text{cov}(X_t, X_{t-2}) \tag{A-4}$$

$$= \text{cov}(\phi X_{t-1} + Z_t, X_{t-2})$$

$$= \text{cov}(\phi^2 X_{t-2} + \phi Z_{t-1} + Z_t, X_{t-2})$$

Due to the linearity property of covariances, we get:

$$\gamma(2) = \phi^2 \operatorname{cov}(X_{t-2}, X_{t-2}) + \phi \operatorname{cov}(Z_{t-1}, X_{t-2}) + \operatorname{cov}(Z_t, X_{t-2})$$

Since there is no autocorrelation between error distribution and time series $\operatorname{cov}(Z_{t-1}, X_{t-2}) = \operatorname{cov}(Z_t, X_{t-2}) = 0$, we get:

$$\gamma(2) = \phi^2 E(X_{t-2}, X_{t-2}) + 0 + 0$$

$$= \phi^2 \gamma(0) = \phi^2 \sigma^2$$

where, σ^2 is the variance of the time series. As the correlation is not zero, this shows that X_t depends on X_{t-2} through X_{t-1}. Due to the iterative nature of the AR models, there is a chain of dependence. We break the dependence by removing the influence of X_{t-1} from both X_t and X_{t-2} to obtain $X_t - \phi X_{t-1}$ and $X_{t-2} - \phi X_{t-1}$, for which the covariance is zero, that is:

$$\operatorname{cov}(X_t - \phi X_{t-1}, X_{t-2} - \phi X_{t-1}) = \operatorname{cov}(Z_t, X_{t-2} - \phi X_{t-1}) = 0$$

Similarly, we obtain zero covariance for X_t and X_{t-3} after breaking the chain of dependence, that is, removing the dependence of the two variables on X_{t-1} and X_{t-2}, for $X_t - f(X_{t-1}, X_{t-2})$ and $X_{t-3} - f(X_{t-1}, X_{t-2})$ for some function f. Continuing this, we would obtain zero covariance for variables $X_t - f(X_{t-1}, X_{t-2}, \ldots, X_{t-\tau+1})$ and $X_{t-\tau} - f(X_{t-1}, X_{t-2}, \ldots, X_{t-\tau+1})$. Therefore, the only covariance which is nonzero is for X_t and X_{t-1} (nothing in between to break the chain of dependence). These covariances with an appropriate function, f, divided by the variance of the process are the partial autocorrelations. Hence, for an AR(1) process, we would have the PACF at lag 1 equal to $\rho(1)$ and at lags > 1 equal to 0. This, together with the tailing off shape of the ACF, identifies the process.

The Partial Autocorrelation Function (PACF) of a zero-mean stationary time series $\{X_t\}_{t=0,1\ldots}$ is defined as:

$$\phi_{11} = corr(X_1, X_0) = \rho(1)$$

$$\phi_{\tau\tau} = corr(X_\tau - f_{(\tau-1)}, X_0 - f_{(\tau-1)}), \tau \geq 2,$$

where

$$f_{(\tau-1)} = f(X_{\tau-1}, ..., X_1)$$

$\phi_{\tau\tau}$ is the correlation between variables X_t and $X_{t-\tau}$ with the linear effect

$$f(X_{t-1}, ..., X_{t-\tau+1}) = \beta_1 X_{t-1} + ... + \beta_{\tau-1} X_{t-\tau+1}$$

on each variable removed.

YULE- WALKER EQUATIONS

To estimate the parameters of the AR model under study, statistical software uses Yule-Walker equations, whose derivation is explained below (Mathuranthan 2014). Consider a generic AR model

$$x_n + a_1 x_{n-1} + a_2 x_{n-2} + ... + a_N x_{n-N} = w_n \quad \text{(A-5)}$$

where N is the order of the model, a_N are the parameters of the model, and w_n is the white noise. Eq. A-5 can be written in compact form as:

$$\sum_{k=0}^{N} a_k x_{n-k} = w_n, a_0 = 1 \quad \text{(A-6)}$$

Multiplying Eq. A-6 by x_{n-l} and taking expectations on both sides, we have:

$$\sum_{k=0}^{N} a_k E(x_{n-k} x_{n-l}) = E(w_n x_{n-l}), a_0 = 1 \quad \text{(A-7)}$$

One can readily identify the auto-correlation and cross-correlation terms in Eq. A-7, which are

$E(x_{n-k}x_{n-l}) = \gamma_{xx}[l-k]$ and $E(w_n x_{n-l}) = \gamma_{WX}[l]$, respectively. Rewriting Eq. A-7, we get

$$\sum_{k=0}^{N} a_k \gamma_{xx}[l-k] = \gamma_{WX}[l] \tag{A-8}$$

The term x_{n-l} can also be obtained from Eq. (A-5) as

$$x_{n-l} = -\sum_{k=1}^{N} a_k x_{n-k-l} + w_{n-l} \tag{A-9}$$

The data and noise are always uncorrelated, therefore $E\{x_{n-k-l}w_n\} = 0$. Also, the autocorrelation of noise is zero at all lags except lag 0, where it is σ^2. These properties are used in the next steps. Restricting lags (k) to only positive values and zero:

$$\begin{aligned}
\gamma_{WX} &= E\{w_n x_{n-l}\} \\
&= E\{w_n(-\sum_{k=1}^{N} a_K x_{n-k-l} + w_{n-l})\} \\
&= E\{-\sum_{k=1}^{N} a_K x_{n-k-l} w_n + w_{n-l} w_n\} \\
&= 0 + E\{w_{n-l} w_n\} \\
&= E\{w_{n-l} w_n\} \\
&= 0, \text{ when } l > 0 \\
&\quad or \\
&= \sigma^2, \text{ when } l = 0
\end{aligned} \tag{A-10}$$

Substituting Eq. A-10 in A-8:

$$\sum_{k=0}^{N} a_K \gamma_{XX}[l-k] = 0, l > 0$$

$$= \sigma^2, l = 0 \tag{A-11}$$

For $l > 0$ case, Eq. A-11 becomes,

$$\sum_{K=1}^{N} a_K \gamma_{XX}[l-k] = -\gamma_{XX}[l] \tag{A-12}$$

Notice the change in the lower limit of summation from $k = 0$ to $k = 1$. Taking the negative sign to the other side and dividing by $\gamma_{XX}[0]$, we get:

$$\sum_{K=1}^{N} a_K \rho_{XX}[k-l] = \rho_{XX}[l] \tag{A-13}$$

Now Eq. A-13 can be written in Matrix form:

$$\begin{matrix} l=1 \\ l=2 \\ \vdots \\ l=N \end{matrix}\begin{bmatrix} \rho_{XX}[0] & \rho_{XX}[1] & \cdots & \rho_{XX}[N-1] \\ \rho_{XX}[-1] & \rho_{XX}[0] & \cdots & \rho_{XX}[N-2] \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{XX}[1-N] & \rho_{XX}[1-N] & \cdots & \rho_{XX}[0] \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} \rho_{XX}[1] \\ \rho_{XX}[2] \\ \vdots \\ \rho_{XX}[N] \end{bmatrix} \tag{A-13}$$

where the matrix contains ACF for different lags depending on the order of the AR model. The Yule-Walker equations comprise a set of N linear equations and N unknown parameters. l is the order of the autoregressive model, which can be varied from 1 to N. For each l, a_N (last term) represents the PACF of order l. For example, for l=1, PACF (1) = a_N, for l=2, PACF (2) = a_N, and so on.

Representing equation (A-13) in a compact form

$$\bar{R}\bar{a} = -\bar{r} \tag{A-14}$$

The solution $\bar{a}$ can be solved by:

$$\bar{a} = -\bar{R}^{-1}\bar{r} \tag{A-15}$$

Once we solve for $\bar{a}$, equivalently a_K, we can find PACF for different orders by determining the last term a_N.

APPENDIX B

In this section we detail the AR (1) models observed for 100 wells from our case study 2. The AR (1) coefficient, along with variance of the white noise, was documented for 40, 50, and 60 months of production data for each well. An interesting observation was that the AR models (including the parameter and the variance) were the same for all the wells, corresponding to the amount of production data used. The models were the same for both OLS and WLS regressions as well. This might be because we had determined that the AR(1) model was best suited for all the wells based on the ACF and PACF of the residuals. Table B.1 summarizes the AR(1) model observed for all 100 wells.

Table B.1: Realized AR (1) model for WLS and OLS residual datasets for 100 Barnett Shale wells.

AR (1) Model Coefficient (OLS and WLS)	Known Production History		
	40 months	50 months	60 months
θ	0.76	0.7469	0.7365
Variance of ε_t	942414	833158	753008

Another observation is that both the coefficients of the model and the variance of the white noise decreased in magnitude with increasing amount of residual data used. This suggests a correlation between AR coefficients and white noise variance with the amount of time series data used, although it is not the scope of this work to determine such a relation. Such an observation is interesting and can be further studied to determine if such a correlation exists and can be quantified.

APPENDIX C

Detailed below is the R code for quantifying uncertainty for a well's gas production forecasts. Each block of code is explained in comments before it.

```
##########################################################################

### Inputting production data from a CSV table and plotting it. The table contains 101 months of production data.

##########################################################################

 RATE<- read.csv('C:/Users/Kishan Joshi/Desktop/Rate data.csv',header=T)[,1]

 time<-1:101

 plot(time, RATE, type='l', xlab="months", ylab="Gas Production, Mscf")

##########################################################################

### If the production data has many zero values, it might give erroneous results. The code below helps to smooth the data, if required.

##########################################################################

quantile(RATE,probs=c(0.05,0.1,0.25))

RATE[RATE<3000]<-NA

RATE<-na.approx(RATE)

plot(time, RATE, type='l', xlab="months", ylab="Gas Production, Mscf")

##########################################################################

### 50 months of production data is used for history matching.

##########################################################################
```

```
t2<- 2:50

RATE.t2<- data.frame(t2=t2,RATE2=RATE[2:50])

parstart<- list(K=10000, n=0.1, a=1)

##############################################################################

### OLS regression is conducted to determine the parameters of LGA model based on 50 months of data.

##############################################################################

fit2<-nlsLM(RATE2~K*a*n*t2^(n-1)/(a+t2^n)^2,start=parstart,data = RATE.t2,trace=T)

q_pred<-predict(fit2,newdata=list(t2=time))

plot(RATE,type='l')

lines(time, q_pred, col="blue")

##############################################################################

### Extracting the residual for time series analysis (OLS)

##############################################################################

Residual<- q_pred-RATE[1:101]

plot(Residual,type='l')

plot(Residual[2:50],type='l')

Error<-ts(Residual1[2:50],frequency=1)

plot(Error)
```

```
##########################################################################

### Determining the AR model to be fitted. Fitting the AR model. Forecasting the residual for the required forecast period. Adding LGA forecast with residual forecast for OLS regression.

##########################################################################

acf(Error,lag.max=49)

pacf(Error,lag.max=49)

arima(Error,order=c(1,0,0))

Error.arima<-arima(Error,order=c(1,0,0))

Error.pred<-forecast.Arima(Error.arima,h=51)

Forecast<-q_pred[51:101]-Error.pred$mean

P10<-q_pred[51:101]-Error.pred$lower[,2]

P90<-q_pred[51:101]-Error.pred$upper[,2]

NWmean<-Forecast

NWP10<-P10

NWP90<-P90

##########################################################################

### WLS regression is conducted to determine the parameters of LGA model based on 50 months of data.

##########################################################################

fit2<-nlsLM(RATE2~K*a*n*t2^(n-1)/(a+t2^n)^2,start=parstart,data=RATE.t2,trace=T,weights=wfct(1/RATE2))
```

```
q_pred<-predict(fit2,newdata=list(t2=time))

plot(RATE,type='l')

lines(time, q_pred, col="blue")

#############################################################################

### Extracting the residual for time series analysis (WLS)

#############################################################################

Residual<- q_pred-RATE[1:101]

plot(Residual,type='l')

plot(Residual[2:50],type='l')

Error<-ts(Residual1[2:50],frequency=1)

plot(Error)

#############################################################################

### Determining the AR model to be fitted. Fitting the AR model. Forecasting the residual for the
required forecast period. Adding LGA forecast with residual forecast for WLS regression.

#############################################################################

acf(Error,lag.max=49)

pacf(Error,lag.max=49)

arima(Error,order=c(1,0,0))

Error.arima<-arima(Error,order=c(1,0,0))

Error.pred<-forecast.Arima(Error.arima,h=51)
```

```
Forecast<-q_pred[51:101]-Error.pred$mean

P10<-q_pred[51:101]-Error.pred$lower[,2]

P90<-q_pred[51:101]-Error.pred$upper[,2]

Wmean<-Forecast

WP10<-P10

WP90<-P90

##########################################################################

### Taking mean of WLS and OLS forecasts. Final P10 is chosen based on the highest P10 from the two regression schemes. Similarly, final P90 is chosen based on the lowest P90 from both the regression schemes.

##########################################################################

Mean<-(Wmean+NWmean)/2

plot(time,RATE, xlab="months", ylab="Gas Production, Mscf",ylim=c(0,60000), type='p')

lines(time[2:50], q_pred[2:50], col="blue",lwd=2)

lines(time[51:101],NWmean, col="red",lwd=2)

lines(time[51:101],NWP10, col="green",type='l',lty=2,lwd=2)

lines(time[51:101],NWP90, col="green",type='l',lty=2,lwd=2)

Fiftyi<- rbind(time,RATE,q_pred)

Fiftyf<-rbind(Mean,NWP10,WP90)

write.csv(Fiftyi,"C:/Users/Kishan Joshi/Desktop/Values/Fiftyi.csv")

write.csv(Fiftyf,"C:/Users/Kishan Joshi/Desktop/Values/Fiftyf.csv")
```